



ARL-TR-10475 • OCT 2026



An Integrated Technology System to Streamline the Military Decision-Making Process for Scalable Cross-Echelon Command and Control (SC3) of the Future (Higher Echelon FY25 Soldier Touchpoint Report)

by Gregory M. Gremillion, Jonroy D. Canady, Julia L. Wright, Angelique A. Scharine, Andrea S. Krausman, Robert Smith, Adam Wachsman, Alex Stauff, William Gerichs, Rhyse Bendell, Jessica Williams, Murat Kucukosmanoglu, Vinicius G. Goecks, Nicholas R. Waytowich, Damon M. Conover, Jeffrey T. Hansberger, Stephen M. Fiore, Justin Brooks, Stephen Gordon, Shuowen Hu, and Amar R. Marathe

DISTRIBUTION STATEMENT A. Approved for public release: distribution unlimited.

NOTICES

Disclaimers

The findings in this report are not to be construed as an official Department of the Army position unless so designated by other authorized documents.

Citation of manufacturer's or trade names does not constitute an official endorsement or approval of the use thereof.

Destroy this report when it is no longer needed. Do not return it to the originator.



An Integrated Technology System to Streamline the Military Decision-Making Process for Scalable Cross-Echelon Command and Control (SC3) of the Future (Higher Echelon FY25 Soldier Touchpoint Report)

**Gregory M. Gremillion, Jonroy D. Canady, Julia L. Wright,
Angelique A. Scharine, Andrea S. Krausman, Nicholas R. Waytowich,
Damon M. Conover, Jeffrey T. Hansberger, Shuowen Hu, and
Amar R. Marathe**

DEVCOM Army Research Laboratory

**Robert Smith, Adam Wachsman, Alex Stauff, William Gerichs, and
Stephen Gordon**

DCS Corporation

Rhyse Bendell, Jessica Williams, and Stephen M. Fiore

University of Central Florida

Murat Kucukosmanoglu and Justin Brooks

D-Prime LLC

Vinicius G. Goecks

Oak Ridge Associated Universities

REPORT DOCUMENTATION PAGE

1. REPORT DATE		2. REPORT TYPE		3. DATES COVERED	
October 2026		Technical Report		START DATE 1 October 2024	END DATE 30 September 2025
4. TITLE AND SUBTITLE An Integrated Technology System to Streamline the Military Decision-Making Process for Scalable Cross-Echelon Command and Control (SC3) of the Future (Higher Echelon FY25 Soldier Touchpoint Report)					
5a. CONTRACT NUMBER		5b. GRANT NUMBER		5c. PROGRAM ELEMENT NUMBER	
5d. PROJECT NUMBER		5e. TASK NUMBER		5f. WORK UNIT NUMBER	
6. AUTHOR(S) Gregory M. Gremillion, Jonroy D. Canady, Julia L. Wright, Angelique A. Scharine, Andrea S. Krausman, Robert Smith, Adam Wachsmann, Alex Stauff, William Gerichs, Rhyse Bendell, Jessica Williams, Murat Kucukosmanoglu, Vinicius G. Goecks, Nicholas R. Waytowich, Damon M. Conover, Jeffrey T. Hansberger, Stephen M. Fiore, Justin Brooks, Stephen Gordon, Shuowen Hu, and Amar R. Marathe					
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) DEVCOM Army Research Laboratory ATTN: FCDD-RLA-IE Aberdeen Proving Ground, MD 21005				8. PERFORMING ORGANIZATION REPORT NUMBER ARL-TR-10475	
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)			10. SPONSOR/MONITOR'S ACRONYM(S)		11. SPONSOR/MONITOR'S REPORT NUMBER(S)
12. DISTRIBUTION/AVAILABILITY STATEMENT DISTRIBUTION STATEMENT A. Approved for public release: distribution unlimited.					
13. SUPPLEMENTARY NOTES ORCIDs: Gregory M. Gremillion, 0000-0002-0205-688X; Jonroy D. Canady, 0000-0003-3555-8779; Julia L. Wright, 0000-0003-3026-1538; Angelique A. Scharine, 0000-0001-8230-1824; Andrea S. Krausman, 0000-0003-1955-8867; Rhyse Bendell, 0000-0003-1988-2528; Jessica Williams, 0000-0002-6056-0312; Murat Kucukosmanoglu, 0000-0003-3611-7692; Vinicius G. Goecks, 0000-0002-4481-671X; Nicholas R. Waytowich, 0000-0002-3786-0675; Stephen M. Fiore, 0000-0003-3529-1322; Amar R. Marathe, 0000-0002-6253-6105					
14. ABSTRACT The Scalable Cross-Echelon Command and Control (SC3) Program Higher Echelon FY25 Soldier Touchpoint evaluated the Integrated Technology System (ITS), a platform comprising three components: the Course of Action (COA)-Generative Pretrained Transformer, a large language model AI; the SC3 Sim, a constructive simulation engine; and the SC3 User Interface, a browser-based GUI. The ITS will assist future planners within distributed, mobile command and control nodes in developing, forecasting, and refining COAs at greater speeds than can be accomplished with existing approaches. Soldier subject matter experts (SMEs) with prior military planning experience from the U.S. Army Mission Command Battle Laboratory (one active-duty Army lieutenant colonel and two retired officers—a former colonel and a former captain) participated in the touchpoint in the roles of brigade-level planning staff. The touchpoint comprised three structured planning sessions with the system, several structured after-action reviews, and several collaborative design discussions. SMEs reported that, even in its current prototype form, the ITS demonstrated strong potential to streamline planning workflows. They estimated that when existing capabilities are more mature and fully integrated into a unified workflow, the ITS could reduce the time and cognitive effort required for COA development by approximately 30% relative to current processes.					
15. SUBJECT TERMS Humans in Complex Systems, military decision-making process, command and control, course of action development, wargaming, large language models, constructive simulation, user interface, teaming, planning					
16. SECURITY CLASSIFICATION OF:				17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES
a. REPORT UNCLASSIFIED	b. ABSTRACT UNCLASSIFIED	c. THIS PAGE UNCLASSIFIED	UU		182
19a. NAME OF RESPONSIBLE PERSON Gregory M. Gremillion				19b. PHONE NUMBER (Include area code) (520) 691-4802	

STANDARD FORM 298 (REV. 5/2020)

Prescribed by ANSI Std. Z39.18

Contents

List of Figures	vii
List of Tables	ix
Executive Summary	x
1. Introduction	1
1.1 Objectives	1
1.2 Methods	1
1.2.1 Integrated Technology System (ITS)	1
1.2.2 Soldier SME Participants	7
1.2.3 Touchpoint Procedures	8
1.2.4 Operational Scenario	11
2. High-Level Findings	13
2.1 Summary of ITS Feedback	14
2.2 Future Research	17
2.3 Most and Least Valuable Features of the ITS	19
2.3.1 What ITS Features Were Useful and How?	20
2.3.2 What Are the Three Most Significant Strengths of COA-GPT Identified by Users?	21
2.3.3 What ITS Features Were Problematic and How?	22
2.3.4 Which Element of the ITS is Most Lacking, Specifically in Terms of Hindering Your Ability to Conduct Planning in a Sufficiently Realistic Manner?	23
2.3.5 What Are the Three Most Significant Areas of COA-GPT for Improvement Identified by Users?	24
2.4 ITS Priority Feature Gaps	25
2.4.1 Temporal Structuring and Phase-Based Simulation	26
2.4.2 Streamlined Tasking and Unit Control	27
2.4.3 Natural Language Editing and COA-GPT Refinement	27
2.4.4 Adversarial and Collaborative Planning Support	28
2.4.5 Additional SME Recommendations	28

2.4.6	Priority Simulation Mechanics Design Goals	29
2.5	ITS Usage Characteristics	30
2.5.1	What Impact Do You Think ITS Had on Your Planning Process?	32
2.5.2	How Satisfied Are Users with COA-GPTs Performance in Facilitating MDMP?	33
2.5.3	How Do Users Compare the Speed and Usefulness of SC3 Sim's Feedback with Traditional Methods of Estimating COA Outcomes?	34
2.5.4	How Well Can SMEs Identify and Trace Back Mission Analysis Outputs to Original Source Documents (e.g., Commander's Intent)? How Effective Is the Interface in Helping Users Access, Trace, and Interpret Mission Analysis Elements Derived from Input Documents?	35
2.5.5	Does the SC3 Interface Clearly Present Essential Mission Analysis Elements (e.g., Constraints, Specified/Essential Tasks, Terrain Limitations) in a Way that Supports Early-Stage Planning?	36
2.5.6	Did interacting with ITS Clarify Any Mission Planning Considerations?	36
2.5.7	How Frequently Do SMEs Revise COA-GPT's Outputs, and What Specific Deficiencies Motivate Those Changes?	37
2.6	Experimental Design Lessons	37
2.6.1	MDMP Individual Studies	38
2.6.2	MDMP Team Studies	39
2.7	Summarized Self-Report Instrument Scores	41
2.7.1	Intake	41
2.7.2	Session 1	41
2.7.3	Session 2	42
2.7.4	Session 3	43
3.	Detailed Analysis and Results	44
3.1	Usability and Features	44
3.1.1	High-Level ITS Feedback	44
3.1.2	Useful Features	45
3.1.3	Problematic Features	47
3.1.4	Lacking Features	49
3.1.5	Planning Effects	50

3.1.6	COA Branches, What Ifs, and Comparisons	54
3.1.7	Trust and Reliance	56
3.2	Team Performance	58
3.2.1	Communication and Awareness	58
3.2.2	Technology Teammate Perception	60
3.2.3	Process and Coordination	61
3.2.4	Cognitive Load	62
3.2.5	Decision-Making	63
3.3	Design Feedback	65
3.3.1	Temporal Structuring and Phase-Based Simulation	65
3.3.2	Streamlined Tasking and Unit Control	66
3.3.3	Natural Language Editing and COA-GPT Refinement	68
3.3.4	Adversarial and Collaborative Planning Support	69
3.3.5	Additional SME Recommendations	70
3.3.6	Priority Simulation Mechanics Design Goals	72
3.4	COA Design Quality Metrics	79
3.4.1	Session 1	80
3.4.2	Session 2	82
3.4.3	Session 3	83
3.5	Interface Interaction Metrics	85
3.5.1	Session 1	86
3.5.2	Session 2	89
3.5.3	Session 3	92
3.6	Cursor Movement Metrics	95
3.6.1	Session 1	96
3.6.2	Session 2	98
3.6.3	Session 3	100
3.7	Eye-Tracking Metrics	103
4.	Conclusion	105
5.	References	107
	Appendix. Supporting Material	109
A.1.	Supporting Materials	110

A.1.1 Usability and Features	110
A.1.1.1 High-level ITS Feedback Cluster	110
A.1.1.2 Useful Features	112
A.1.1.3 Problematic Features	114
A.1.1.4 Lacking Features	115
A.1.1.5 Planning Effects	116
A.1.1.6 Trust and Reliance	122
A.1.2 Team Performance	132
A.1.3 Design Feedback	140
A.2. Self-Report Instruments	146
A.2.1 User Traits and Individual Differences	146
A.2.1.1 Demographics	146
A.2.1.2 Rational and Intuitive Decision Style Scale	146
A.2.1.3 Intolerance of Uncertainty Scale	147
A.2.1.4 Propensity to Trust (LLM Agents)	149
A.2.1.5 Team Process Scale (Short)	149
A.2.2 Workload	151
A.2.2.1 NASA-TLX (Modified)	151
A.2.2.2 Team Workload Questionnaire	152
A.2.3 System Evaluations	153
A.2.3.1 Human-Computer Trust (Short)	153
A.2.3.2 User Interface Satisfaction	155
A.2.3.3 Technology Acceptance Model	156
A.2.3.4 System Trustworthiness Scale	157
A.2.4 Touchpoint Tailored Evaluations	158
A.2.4.1 SC3 Touchpoint Session 1 ITS Evaluation	158
A.2.4.2 SC3 Touchpoint Session 2 ITS Evaluation	159
A.2.4.3 SC3 Touchpoint Session 3 ITS Evaluation	161
A.2.4.4 System Elements Evaluation	163
A.2.4.5 MCBL BOLT Questionnaire (Modified)	164
List of Symbols, Abbreviations, and Acronyms	166
Distribution List	168

List of Figures

Figure 1.	COA-GPT conceptual flowchart.....	2
Figure 2.	COA-GPT UI generate MDMP products page (COA development tab).	3
Figure 3.	SC3 sim combat representation diagram.....	4
Figure 4.	SC3 sim movement representation diagram.....	4
Figure 5.	SC3 UI COA editor page (COA creation tab).....	5
Figure 6.	SC3 UI COA editor page (edit positions tab).	6
Figure 7.	SC3 UI COA analysis page.....	6
Figure 8.	SC3 UI operation overview page.	7
Figure 9.	Operation Tropic Tortoise CONOPS.	12
Figure 10.	Operation Tropic Tortoise enemy COA.....	12
Figure 11.	Session 2 design feature ratings.	26
Figure 12.	Future unit configuration concept.	74
Figure 13.	Future combat area concept.....	75
Figure 14.	Future range effects concept.....	76
Figure 15.	Future indirect fires concept.....	77
Figure 16.	Future terrain occlusion concept.	78
Figure 17.	Future terrain classification concept.	79
Figure 18.	Future movement effects concept.	79
Figure 19.	Session 1 destroyed unit fraction.	80
Figure 20.	Session 1 final combat power.....	81
Figure 21.	Session 2 destroyed unit fraction.	82
Figure 22.	Session 2 final combat power.....	83
Figure 23.	Session 3 destroyed unit fraction.	84
Figure 24.	Session 3 final combat power.....	85
Figure 25.	UI core areas of interest.	86
Figure 26.	Session 1 AOI usage distribution.	87
Figure 27.	Session 1 transition matrix for Alpha.....	88
Figure 28.	Session 1 transition matrix for Echo.	88
Figure 29.	Session 1 transition matrix for Golf.	89
Figure 30.	Session 2 AOI usage distribution.	90
Figure 31.	Session 2 transition matrix for Alpha.....	91

Figure 32.	Session 2 transition matrix for Echo.	91
Figure 33.	Session 2 transition matrix for Golf.	92
Figure 34.	Session 3 AOI usage distribution.	93
Figure 35.	Session 3 transition matrix for Alpha.....	94
Figure 36.	Session 3 transition matrix for Echo.	94
Figure 37.	Session 3 transition matrix for Golf.	95
Figure 38.	Session 1 cursor position heatmaps.....	97
Figure 39.	Session 1 cursor straightness distributions.....	98
Figure 40.	Session 2 cursor position heatmaps.....	99
Figure 41.	Session 2 cursor straightness distributions.....	100
Figure 42.	Session 3 cursor position heatmaps.....	101
Figure 43.	Session 3 cursor straightness distributions.....	102
Figure 44.	Select eye features.	104
Figure A-1.	Session 2 ITS evaluation.	117
Figure A-2.	Session 1 COA-GPT evaluation.....	118
Figure A-3.	Session 2 COA-GPT evaluation.....	120
Figure A-4.	Session 2 ITS evaluation.	120
Figure A-5.	Session 1 human-computer trust scale (short).....	122
Figure A-6.	Session 2 human-computer trust scale (short).....	124
Figure A-7.	Session 2 system trustworthiness scale.	126
Figure A-8.	Session 3 ITS evaluation.	129
Figure A-9.	Session 3 COA-GPT & simulator evaluation.	129
Figure A-10.	Session 3 human-computer trust (COA-GPT).	131
Figure A-11.	Session 2 ITS evaluation - team.	136
Figure A-12.	Session 2 team process scale.	138
Figure A-13.	Session 1 questionnaire for user interface satisfaction – overall (items 1-6).	140
Figure A-14.	Session 1 questionnaire for user interface satisfaction – features (items 7-28).	141
Figure A-15.	Session 2 technology acceptance model.	143
Figure A-16.	Session 2 design feature ratings. Responses for SME Alpha were not captured for this questionnaire.	144

List of Tables

Table 1.	Summary of scored intake questionnaire responses.....	41
Table 2.	Summary of scored session 1 questionnaire responses.	42
Table 3.	Summary of scored session 2 questionnaire responses.	43
Table 4.	Summary of scored session 3 questionnaire responses.	44
Table 5.	Session 2 team process scale summary.	64
Table 6.	Eye-tracking features extracted.	103
Table A-1.	Session 2 ITS evaluation.	118
Table A-2.	Session 1 COA-GPT evaluation.....	119
Table A-3.	Session 2 COA-GPT evaluation.....	120
Table A-4.	Session 2 ITS evaluation.	121
Table A-5.	Session 1 human-computer trust scale (short).....	123
Table A-6.	Session 2 human-computer trust scale (short).....	125
Table A-7.	Session 2 system trustworthiness scale.	126
Table A-8.	Session 3 ITS evaluation.	129
Table A-9.	Session 3 COA-GPT & simulator evaluation.	130
Table A-10.	Session 3 human-computer trust (COA-GPT).	132
Table A-11.	Session 2 ITS evaluation - team.	137
Table A-12.	Session 2 team process scale.	139
Table A-13.	Session 1 questionnaire for user interface satisfaction – overall (items 1-6).	140
Table A-14.	Session 1 questionnaire for user interface satisfaction – features (items 7-28).	142
Table A-15.	Session 2 technology acceptance model.	144
Table A-16.	Session 2 design feature ratings.	145

Executive Summary

The Scalable Cross-Echelon Command and Control (SC3) Program Higher Echelon FY25 Soldier Touchpoint evaluated the Integrated Technology System (ITS), a planning and decision-support platform that includes three primary components: the Course of Action (COA)-Generative Pretrained Transformer (GPT) large language model (LLM)-based AI planning tool, the SC3 Sim constructive simulation engine, and the SC3 User Interface browser-based GUI. The ITS is intended to assist military planners in developing, forecasting, and refining COAs, both individually and within a team, as part of future distributed, mobile command and control (C2) nodes at greater speeds than can be accomplished with existing traditional, and even state-of-the-art, approaches.

- COA-GPT: A generative AI system that leverages LLMs and a corpus of doctrinal and operation-specific documentation to conduct the first several steps of the Military Decision-Making Process (MDMP) to produce various planning products and a set of output seed COAs that represent distinct initial plans, on the order of minutes.
- SC3 Sim: A lightweight constructive simulation engine that rolls out a set of domain-relevant movement and combat mechanics for a multi-day, operational scale mission as specified by an input COA to produce various outcome trajectories and performance metrics for the simulated aggregate units, on the order of seconds.
- SC3 UI: A browser-based GUI that allows multiple users to query COA-GPT for seed COAs, simulate them within SC3 Sim, review their outcomes, and manipulate their configuration in a rapid, user-friendly manner.

Three subject matter experts (SMEs) from the U.S. Army Mission Command Battle Laboratory, comprising one active-duty U.S. Army lieutenant colonel and two retired former Soldiers (a former colonel and a former captain), participated in the touchpoint. All had prior MDMP experience. The touchpoint comprised three structured planning sessions using the ITS in the role of planning staff at a brigade level, and several structured after-action reviews and collaborative design discussions. During the three planning sessions, they assessed the current capabilities of the ITS and provided detailed feedback on its future development priorities and operational applicability. The first planning session consisted of individual planning as separate ITS stations, where they did not collaborate nor have access to use COA-GPT to modify the seed COAs. The second planning session consisted of collective planning at a single ITS station, with periods of break-out individual work at separate ITS stations, where they did not have access to use COA-GPT to modify the seed COAs. The third planning session consisted

of both individual and collective planning, with periods of break-out individual work at separate ITS stations and group convergence around a single ITS station, where they did have access to use COA-GPT to modify the seed COAs. This touchpoint represents the first milestone test of this prototype platform, which is in its earliest state of minimum viable capability.

SMEs reported that, even in its early and partially integrated form, the ITS demonstrated meaningful promise. They consistently expressed enthusiasm for the system's direction and indicated that its current features already show strong potential to streamline planning workflows. Specifically, SMEs estimated that once the existing capabilities (such as initial seed COA generation, structured unit task assignment, iterative plan editing, and simulated COA results) are more mature and fully integrated into a unified workflow, the system could reduce the time and cognitive effort required for COA development by approximately 30%, relative to planning processes with which they have prior experience.

SMEs highlighted that this functionality would be especially valuable for individual users operating under time pressure, and that the ability to start with a generated COA and make structured, testable edits marked a major improvement over current, manual methods. They strongly supported the completion of these core features, describing it as a meaningful short-term milestone that would deliver real value while also laying the groundwork for future expansion.

Looking ahead, SMEs described a clear and compelling vision for a goal-state ITS that could support collaborative planning at scale. They suggested that a fully realized version of the system could reduce overall planning cost and effort by 60%–70%, provided it includes additional capabilities to support distributed teams. These features include the ability to conduct phase-based planning, wargame against a human adversary, assign tasking and geographic ownership by role, and support traceable COA development across all stages of planning. In their view, the ITS has potential not only to accelerate planning, but to fundamentally improve how planning teams share understanding, coordinate responsibilities, and adapt to changing mission demands.

SMEs also emphasized that realizing this potential will require testing the ITS rigorously and in realistic planning settings. We recommend that this will require structured, comparative experimentation designed to test the ITS and compare it directly with traditional methods. Future research must begin with the establishment of valid baselines using current planning methods, followed by matched scenarios that test the same types of missions using ITS-supported workflows. To capture the full impact of the system, scenarios should incorporate a spectrum of combat- and resource-focused planning tasks and introduce

controlled perturbations such as updated commander's intent, new constraints, or evolving operational intelligence. Equally important is ensuring that users are trained not just in the system but in their specific planning roles so that evaluation reflects genuine team performance. This experimentation must be designed to support cumulative findings and long-term learning, rather than disconnected, one-off demonstrations.

Overall, the touchpoint confirmed that the ITS is on a strong development path. SMEs responded positively to the system's current capabilities, emphasized its immediate utility for individual users, and outlined a detailed and achievable vision for expanding its value in team-based planning environments. With focused development of its core functions and a deliberate, structured research agenda, the ITS has the potential to significantly improve how military teams plan, adapt, and coordinate in complex operational contexts.

1. Introduction

This report details the Scalable Cross-Echelon Command and Control (SC3) Higher Echelon (HE) FY25 Soldier Touchpoint (STP), an event conducted by a team of researchers and engineers from the U.S. Army Combat Capabilities Development Command (DEVCOM) Army Research Laboratory (ARL), DCS Corporation, the University of Central Florida (UCF), and D-Prime LLC, in collaboration with the U.S. Army Mission Command Battle Laboratory (MCBL). The objective of the event was to test a novel suite of integrated technologies the team developed for dramatically accelerated command and control (C2) mission planning.

1.1 Objectives

The primary aim of the SC3 program is to develop and test materiel and nonmaterial solutions to address C2 challenges and needs in the future operational environment (FOE), which include the following: 1) accelerating the operations process to identify, create, and exploit windows of opportunity against an adversary; 2) incorporating novel technologies (e.g., AI) into the operations process while maintaining transparency, trust, and unity of effort; 3) conducting the operations process within a constellation of mobile, distributed C2 nodes; and 4) maintaining C2 overmatch against a technology-enabled near-peer adversary. The touchpoint and the development of technologies it examines is a component of the SC3 program led by a team in ARL's Humans in Complex Systems (HCxS) division that is conducting early applied research (6.2 level) focused on novel C2 capabilities at HEs, which we define as brigade and above. The HCxS 6.2 HE team (itself comprising personnel from ARL, DCS, UCF, and D-Prime) was supported by collaborators from the HCxS 6.2 lower echelon team, the HCxS basic research (6.1 level) team, and the Military Information Sciences division of ARL. The goal of the touchpoint is to capture data on, assess, and explore 1) the impacts of an integrated technology system for rapid mission planning on portions of the Military Decision-Making Process (MDMP) when employed by Soldier subject matter expert (SME) users, 2) areas of design and implementation improvement, and 3) approaches to improve future experimentation with technology-enabled C2. This serves as the first test of these integrated prototype technologies.

1.2 Methods

1.2.1 Integrated Technology System (ITS)

Developed by the SC3 HCxS 6.2 HE team, the ITS is a planning and decision-support platform intended to assist military personnel in developing, forecasting,

and refining courses of action (COAs), both individually and as part of distributed planning teams at greater speeds than can be accomplished with existing traditional, and even state-of-the-art, approaches. The ITS comprises three primary components: the COA-Generative Pretrained Transformer (GPT) large language model (LLM)-based AI system, the SC3 Sim constructive simulation engine, and the SC3 User Interface (UI) browser-based GUI.

1.2.1.1 COA-GPT

COA-GPT is an LLM AI tool that generates output products for the first three steps of MDMP: receipt of mission, mission analysis, and COA development (Headquarters, Department of the Army 2024). This system leverages state-of-the-art foundation multimodal LLMs, chain-of-thought prompt engineering approaches crafted in collaboration with military experts, and retrieval-augmented generation of military doctrine source documents (Figure 1). The version used in this touchpoint significantly improves on prior versions developed by the basic research (6.1 level) collaborators within the SC3 program (Goecks and Waytowich 2024). The COAs that are produced by COA-GPT at the end of its COA development step are communicated over a local network to achieve a loose, unidirectional integration with the other elements of the ITS. Users can specify the number of distinct COAs it attempts to generate. COA-GPT also has a dedicated UI (Figure 2) that was used in tandem with the ITS UI (SC3 UI) during the last of the touchpoint planning sessions.

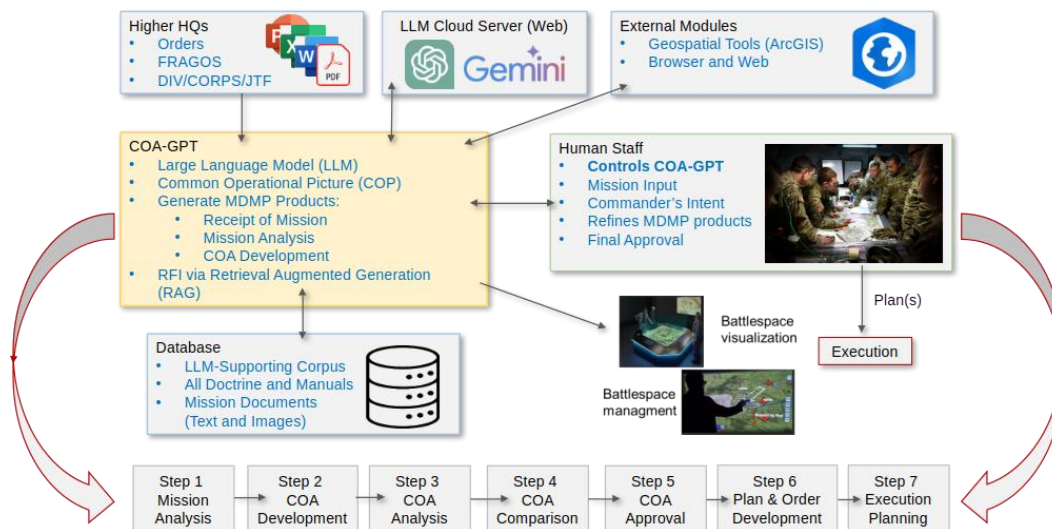


Figure 1. COA-GPT conceptual flowchart.

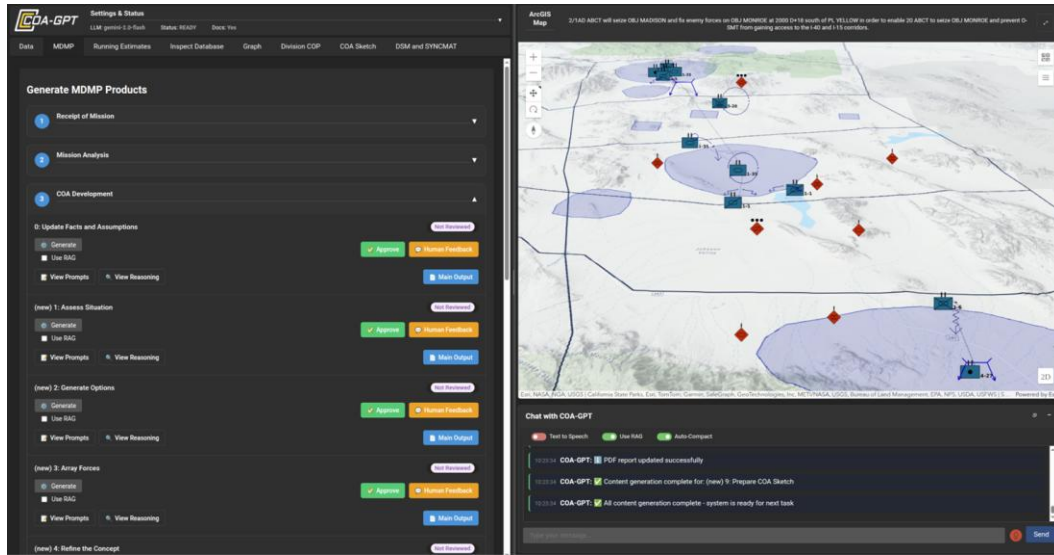


Figure 2. COA-GPT UI generate MDMP products page (COA development tab).

1.2.1.2 SC3 Sim

SC3 Sim is a constructive simulation engine that takes a configuration specifying the COA as an assignment of the friendly and enemy units' tactical mission tasks (using 1-h resolution) as input, then rapidly computes the outcome as dictated by a set of domain-specific mechanics. This simulation platform was developed by the 6.2 HE team with a focus on simulating COA outcomes rapidly in a lightweight, scalable, and extensible architecture. This initial prototype implementation aimed to capture coarse mechanics for combat and movement of aggregated units to support approximate projections of outcomes rather than entity-level simulation or high-fidelity models of behavior and effects, as these were deemed the most fundamental dynamics for assessing a COA. Combat engagement is initiated automatically by a collision algorithm when the notional region of maximum effective range of a unit's primary weapon system (denoted "Combat Area") overlap with the notional region of a target unit's physical location (denoted "Formation Area"), resulting in attrition of the target unit at a rate that is dependent on the combat power of each unit and draws from parameters of a doctrinal correlation of forces and means table (Figure 3). Movement of units is determined by a pathing algorithm on a navigation mesh of traversable terrain as dictated by a simplified modified combined obstacles overlay, indicating non-navigable regions of severely restricted terrain (Figure 4).

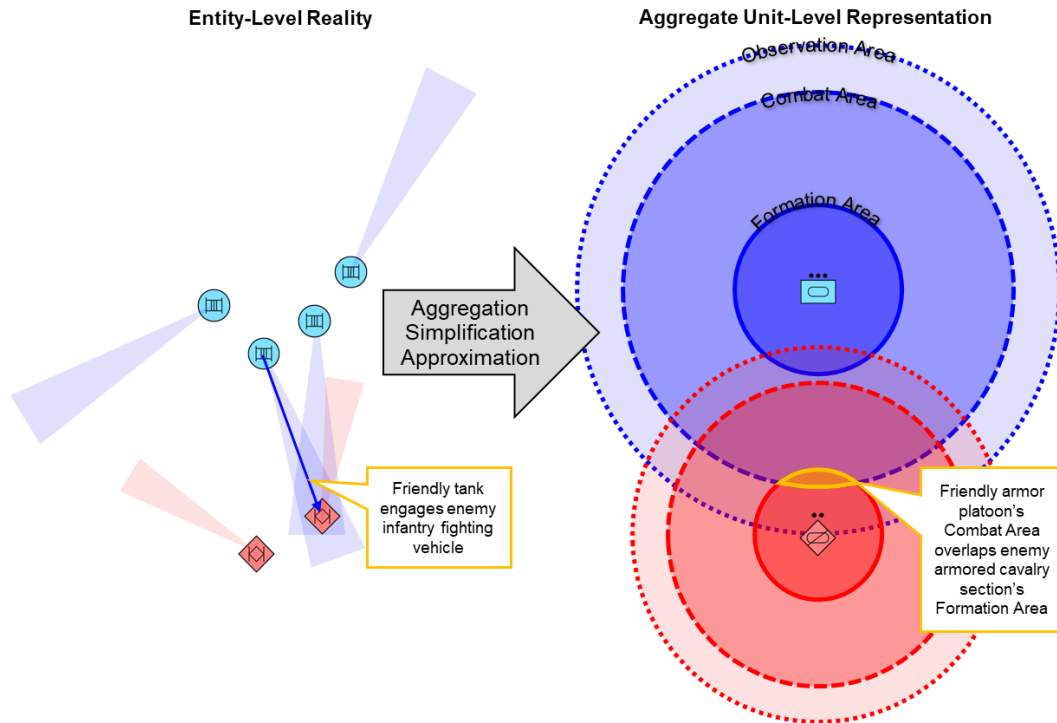


Figure 3. SC3 sim combat representation diagram.

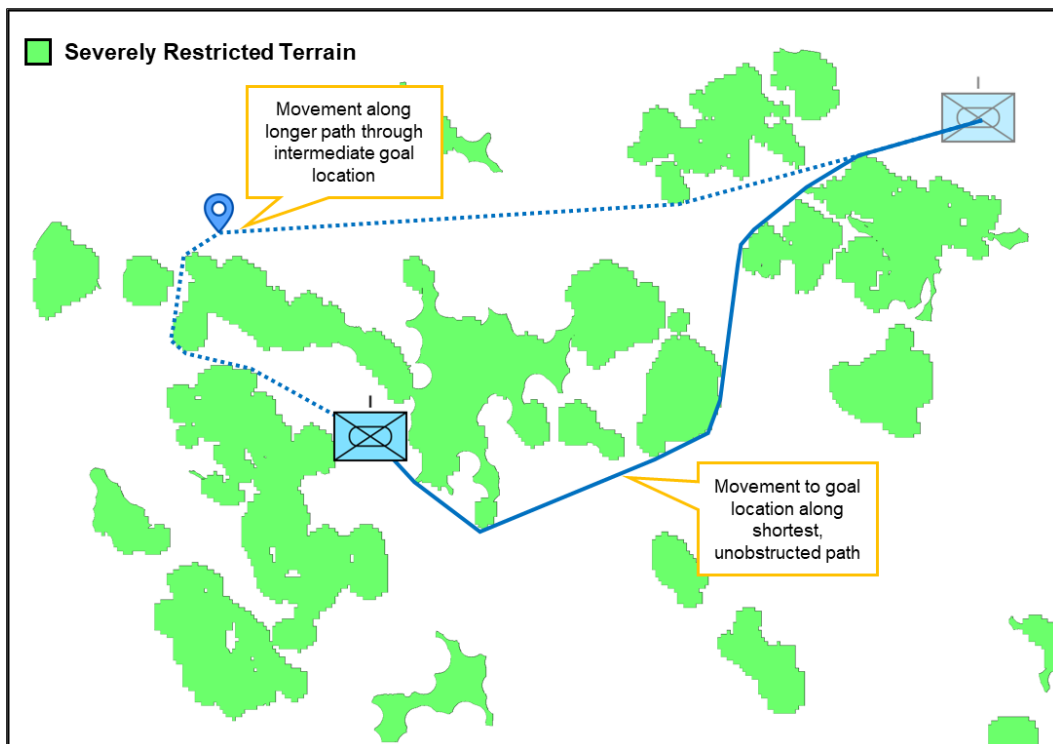


Figure 4. SC3 sim movement representation diagram.

1.2.1.3 SC3 UI

SC3 UI is a browser-based GUI that enables users to interact with SC3 Sim, load seed COA configurations transmitted from COA-GPT, duplicate COAs, manually edit COA configurations, initiate simulation of the configured COA, and view the simulated COA outcomes in spatial and tabular data formats. The aim of developing this UI was to have a minimum viable capability for interacting with the COA-GPT and SC3 Sim to support testing of novel MDMF procedures and concepts within the accessible form factor of computer screen, keyboard, and mouse interface devices. The team is pursuing more expansive, further-term UI concepts in a parallel effort (OpsVision), leveraging technologies for head-mounted, mixed reality, multi-modal, and naturalistic spatial interfaces, which will be informed by the observed usage of, and feedback on, this initial screen-based implementation of the SC3 UI. Figures 5–8 depict the major interface elements of the SC3 UI.

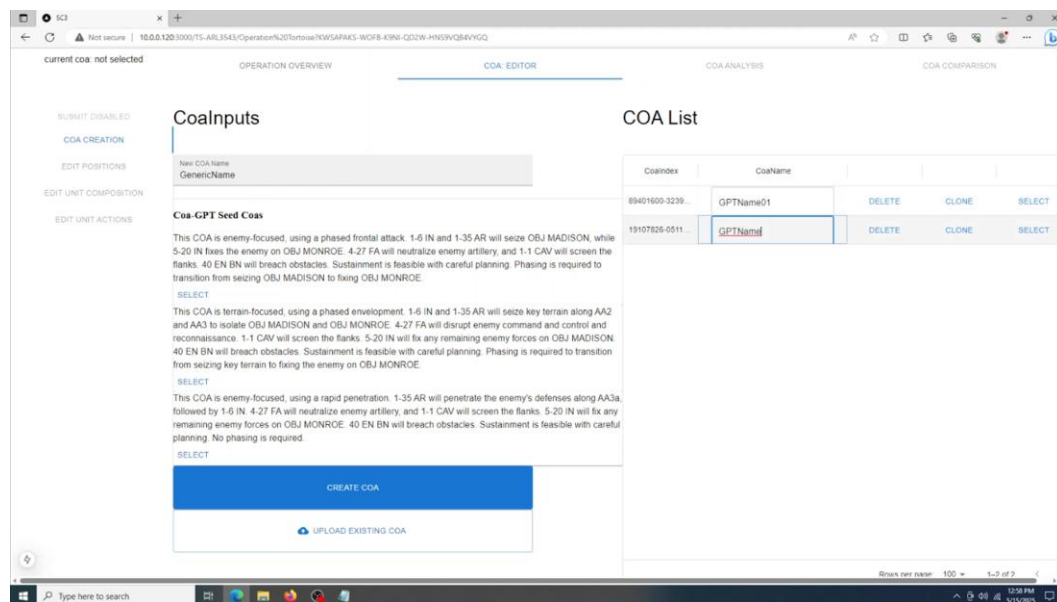


Figure 5. SC3 UI COA editor page (COA creation tab).

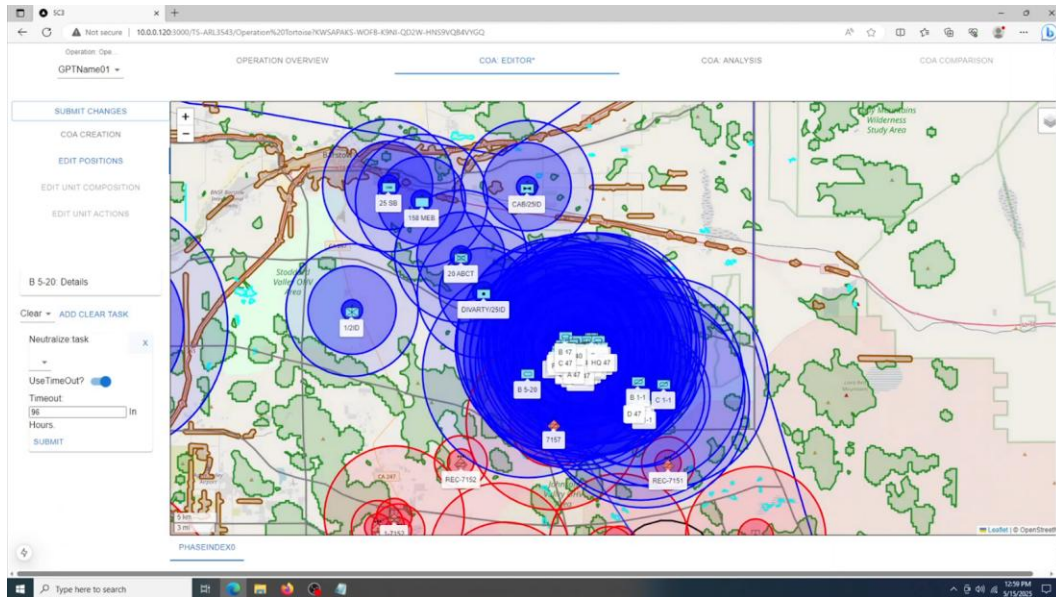


Figure 6. SC3 UI COA editor page (edit positions tab).

Operation: GPTName01

Result ID: 85505502-0750-1703-1403-074927039174 | RealTimeTaken: 11 | SimTimeTaken: 353880

Blue Force: Dead or alive Combat power 1271.23/1357.89 34 / 51				Red Force: Dead or alive Combat power 504.00/600.80 33 / 67			
unitID	FieldName	Combat Power	IsAlive	unitID	FieldName	Combat Power	IsAlive
c0d9891-647-470e-a407-c326...	DIARTY/25ID	150.00 -> 150.00 = 100.00%	Alive	679709eb-0290-4190-8800-01...	8523 852	12.00 -> 12.00 = 100.00%	Alive
7269a110-0d2e-4019-ba4a-901...	25 98	150.00 -> 150.00 = 100.00%	Alive	87aced05-1cef-4832-0629-814c...	2/5-7158	3.00 -> 0.00 = 0.00%	Dead
2c88085e-0e2f-400a-8622-c3a...	1/2ID	150.00 -> 150.00 = 100.00%	Alive	0636907b-2505-4c7c-86ef-692...	1MTR-7151	3.00 -> 0.00 = 0.00%	Dead
0876eab7-e3b5-4e4a-a2b9-64...	158 MEB	150.00 -> 150.00 = 100.00%	Alive	ca1d6c2-6955-4892-8bca-e8b...	1/TA-7158	1.00 -> 1.00 = 100.00%	Alive
694c953-04d3-4091-8c7f-d8c4...	20 ABCT	150.00 -> 150.00 = 100.00%	Alive	dfe42204-c903-4db8-a02c-626...	3-7153	5.00 -> 0.92 = 18.44%	Alive
0f57ab07-86e5-4e4e-a95f-109...	CAB/25ID	150.00 -> 150.00 = 100.00%	Alive	df224705-aaf7-4351-0390-306a...	1MTR-7152	3.00 -> 3.00 = 100.00%	Alive
19e47d54-280a-41c7-023e-842...	HQ 1-1	0.10 -> 0.00 = 0.00%	Dead	7aa83331-5e10-47f6-a932-630...	HQ 715	5.00 -> 5.00 = 100.00%	Alive
0f37853f-6095-47d2-001b-c0e...	A 1-1	8.55 -> 0.00 = 0.00%	Dead	ea70f2f-244b-47e9-a87b-c247...	1-7151	5.00 -> 5.00 = 100.00%	Alive

Blue Force: Tasking			Red Force: Tasking		
unitName	TaskName	TaskStatus	unitName	TaskName	TaskStatus
H 47	task-1-35-AR-1-prepare	Succeeded			
H 47	task-1-35-AR-2-seize	Succeeded			
H 47	0696c2fc-495c-4d38-a5fe-4dc504e77e97	Succeeded			

Figure 7. SC3 UI COA analysis page.

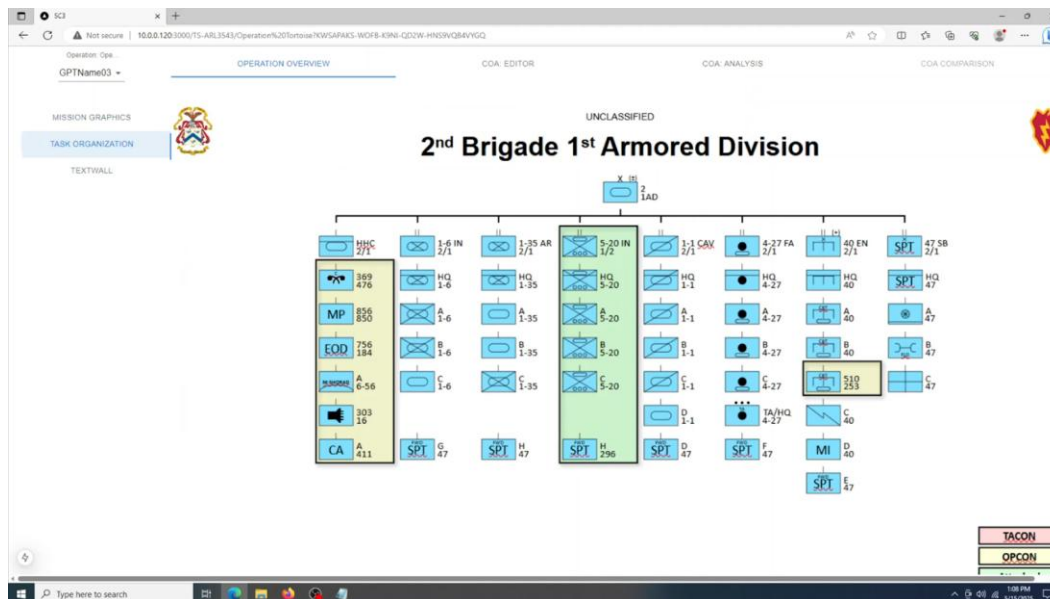


Figure 8. SC3 UI operation overview page.

1.2.2 Soldier SME Participants

Three Soldier SMEs from MCBL (one active-duty U.S. Army lieutenant colonel and two retired former Soldiers—a former colonel (COL) and a former captain) participated in the touchpoint as users of the ITS. All had extensive experience with COA development (ranging from 9 to 20+ years), and all reported they were familiar with AI/ML terms and concepts, but none considered themselves experts. Two SMEs had participated in a prior COA-GPT developmental testing event but were otherwise unfamiliar with COA-GPT at its current level of development. Two of the three SMEs were familiar with the fictional “Operation Tropic Tortoise” scenario that the experimental scenarios were based upon, and they briefed the other SME on the scenario’s COA-relevant details before beginning.* Each SME was assigned an individual phonetic alphabet designator (i.e., Alpha, Echo, and Golf), by which they are referred to throughout this report.

The SMEs completed several intake questionnaires before beginning the touchpoint. The Intolerance of Uncertainty measure (Carleton et al. 2007) assesses two factors, prospective anxiety and inhibitory anxiety, using four-point Likert scales. Higher scores indicate a person may experience anxiety when dealing with

*Operation Tropic Tortoise is scoped for a planning staff within the 2nd Brigade 1st Armored Division, operating as part of an operation of the 25th Infantry Division against a near-peer adversary comprising a Russian Brigade Combat Group in terrain situated in southern California. The scenario documents were originally generated as training material, containing all necessary scenario details to instruct students in MDMP (e.g., operations order, order of battle, maps and terrain overlays, and enemy COAs), and were provided by MCBL. The scenario is described further in Section 0.

ambiguous situations. All SMEs scored below average (2.5) for inhibitory anxiety, indicating it was unlikely anxiety over ambiguity would prevent them from responding in a timely manner. One SME scored highly for prospective anxiety (3.25); however, that person also had over 20 years' experience developing COAs. As such, it was unlikely they would find the exercise anxiety-provoking.

The Rational and Intuitive Decision Style scale (Hamilton et al. 2016) assesses whether a person is more likely to make decisions rationally or intuitively. It comprises two five-item scales that are scored using five-point Likert scales. All SMEs scored below the mean (3.0) for intuitive decisions and above average for rational decision styles.

The Propensity to Trust LLM Agents scale (based on the Propensity to Trust Intelligent Agents scale [Schneider et al. 2017]) measures an individual's attitudes towards technology and, subsequently, the potential for their collaboration with technology. The scale comprises six items scored using five-point Likert scales ($\alpha = 0.84$). The SMEs each scored near the midpoint of this scale (3.0), indicating while they do not report high trust in LLM agents, they also do not have low trust in agents. This suggests they would be well-suited to participate in an exploratory exercise evaluating COA-generation support technology.

1.2.3 Touchpoint Procedures

Following an initial intake activity, which captured data on the SME's individual characteristics, they participated in a series of planning sessions, capturing data on their interactions with, and perceptions of, the ITS. These sessions spanned the three-day study period, and each session was typically more than 1 h in duration. Each planning session explored either a different set of ITS capabilities or a different mode of interaction and planning between the SMEs. This variation was meant to capture informative, but preliminary, findings across a wider range of formats and functions, rather than repeated measures of a fixed test condition. In addition to these planning sessions, the SMEs participated in guided after-action review (AAR) discussion with the research staff, to capture qualitative feedback on their perceptions of the ITS. Lastly, the SMEs participated in several design discussions to capture their insights on how future improvements and refinements to the ITS might be implemented or prioritized.

1.2.3.1 Intake

Subjects responded to a series of self-report instruments prior to completing any of the touchpoint tasks (Appendix, Section A.2). These included the series of instruments to capture user traits and individual differences described above. The subjects also completed initial baseline questionnaires to capture some measure of

the individual and team workload they perceived retrospectively when engaging in traditional MDMP tasks in their past experiences. This was used as a proxy measure due to the absence of a baseline condition of conducting manual MDMP in this touchpoint. The scores for these intake questionnaires are provided in Section 2.7. The self-report instruments administered were as follows:

- Demographics - Appendix, Section A.2.1.1
- Rational and Intuitive Decision Style Scale - Appendix, Section A.2.1.2
- Intolerance of Uncertainty Scale - Appendix, Section A.2.1.3
- Propensity to Trust (LLM Agents) - Appendix, Section A.2.1.4
- NASA Task Load Index (NASA-TLX) (modified) - Appendix, Section A.2.2.1
- Team Workload Questionnaire - Appendix, Section A.2.2.2

1.2.3.2 Session 1

Session 1 involved the execution, evaluation, and alteration of seed COAs by each participant individually. The goals of this session were to 1) familiarize them with usage of the simulation capabilities of the ITS without interacting with COA-GPT and 2) assess a more simplified and stove-piped planning process without any group interaction during planning. Three seed COAs were provided to participants and they did not have the opportunity to generate new COAs using COA-GPT. Participants could access the edit positions, operation overview, and COA analysis tabs, but could not access the COA comparison tab or load the results from more than one COA at once. After completing session 1, the SMEs completed questionnaires assessing individual-based measures of workload, trust in the ITS, and user evaluations of the interface and ITS. The self-report instruments administered were as follows:

- NASA-TLX (modified) - Appendix, Section A.2.2.1
- Human-Computer Trust (Short) - Appendix, Section A.2.3.1
- UI Satisfaction - Appendix, Section A.2.3.2
- SC3 Touchpoint Session 1 ITS Evaluation - Appendix, Section A.2.4.1

1.2.3.3 Session 2

Session 2 involved all three participants working together to develop the best COA that they could in a collaborative fashion. The goal of this session was to assess a more complex and collaborative planning process, assuming the SMEs had gained some proficiency with usage of the simulation capabilities of the ITS. All three computers with the simulator and preloaded seed COAs (different from session 1) were made available to the participants; however, they were told that they would need to come up with a single, best COA through their collaboration. The general

flow of session 2 consisted of an initial decision by the participants to huddle around the middle computer to evaluate the outputs of the initial seed COAs, during which they identified opportunities for alterations, then they split out to their individual computers to make alterations and explore the results of those COA amendments. Next, they engaged in two or three check-ins (not all triadic interactions, two separate dyad discussions), and then concluded with a full team examination of the best COA and avenues for alternatives before determining that they had done the best they could, given the capabilities of the simulator. After completing session 2, the SMEs completed questionnaires assessing individual and team-based measures of workload, trust in the ITS, team processes, perceived trustworthiness, usefulness and ease of use of the ITS, and user evaluations of the interface and ITS. The self-report instruments administered were as follows:

- NASA-TLX (modified) - Appendix, Section A.2.2.1
- Team Workload Questionnaire - Appendix, Section A.2.2.2
- Team Process Scale (Short) - Appendix, Section A.2.1.5
- SC3 Touchpoint Session 2 ITS Evaluation - Appendix, Section A.2.4.2
- System Elements Evaluation - Appendix, Section A.2.4.4
- System Trustworthiness Scale - Appendix, Section A.2.3.4
- Human-Computer Trust (Short) - Appendix, Section A.2.3.1
- Technology Acceptance Model - Appendix, Section A.2.3.3
- MCBL BOLT Questionnaire (Modified) - Appendix, Section A.2.4.5

1.2.3.4 Session 3

Session 3 involved the capability to modify seed COAs through direct interaction with the COA-GPT system. The goal of this session was to assess the end-to-end capabilities of the ITS, providing both the COA generation and simulation capabilities of the ITS, and an even more complex planning process that involved both individual and group planning activities. During this session, participants were instructed to spend a maximum of 10 min discussing the initial seed COAs that were previously generated and provided, and then return to their individual stations and work to refine COAs as they see fit, including usage of COA-GPT to modify the pregenerated COAs. After 1 h, participants were asked to reconvene to discuss their COA development and work together towards a final, best COA. After completing session 3, the SMEs completed questionnaires assessing individual-based measures of workload (NASA-TLX), trust in the ITS, and user evaluations of COA-GPT specifically and ITS as a whole. The self-report instruments administered were as follows:

- NASA-TLX (modified) - Appendix, Section A.2.2.1

- Human-Computer Trust (Short, “COA-GPT” Version) - Appendix, Section A.2.3.1
- SC3 Touchpoint Session 3 ITS Evaluation - Appendix, Section A.2.4.3

1.2.3.5 AAR

Following each planning session, SMEs provided feedback on aspects of their experience working with the ITS. The review was conducted in a panel format during which researchers posed questions and SMEs were encouraged to respond. Generally, questions for additional elaboration were prepared in advance (see Appendix). However, some lines of inquiry arose organically from observations of participants’ activities. For example, feedback regarding the ability to represent and modify geometries on the map interface was collected from follow-up queries based on commentary made by participants during session 1.

1.2.3.6 Design Discussions

Multiple structured discussions were conducted throughout the touchpoint in a format of conversation between collaborators rather than observation of participants, in an effort to capture the SMEs’ input on the usability of the ITS and the direction of its future design and development. These discussions focused on identification of critical gaps of the ITS and approaches to close them, such as desired UI functions and simulation mechanics implementation details.

1.2.4 Operational Scenario

The planning that was performed by the SMEs using the ITS was scoped as though they were part of a brigade-level staff, conducting COA development within a future, technology-enabled MDMP executing the aforementioned fictional Operation Tropic Tortoise scenario (developed for educational use in the Army’s Command General Staff College and provided to the research staff by MCBL). This scenario captures a large-scale combat operation (LSCO) with U.S. forces in an offensive mission against a near-peer adversary situated to the south in a defensive posture. The scenario provides an operations order (OPORD) from 25th Infantry Division (25ID) higher headquarters directing 2nd Brigade Combat Team, 1st Armored Division (2/1AD) to first seize the brigade’s named objective, OBJ MADISON, and then support seizure of the division’s named objective, OBJ MONROE, by fixing the enemy on that objective from the north to enable envelopment by an adjacent friendly unit, 20th Armored Brigade Combat Team. The documentation provided by MCBL describes the scenario in extensive detail, including the following components: friendly unit concept of operations (CONOPS), schemes of maneuver, intelligence, fires, protection, engineering,

sustainment, and order of battle for 25ID and its subordinate units; mission graphics, identified avenues of approach, key terrain, obstacles, cross-mobility corridors, and weather forecasts for the area of interest (AOI) and area of operation (AO); composition, disposition, and expected course of action for the enemy 715th Brigade Combat Group (715 BCG). Examples of the scenario documentation, the 25ID CONOPS, and 715 BCG COA, respectively, are shown in Figures 9 and 10.

Concept of Operations

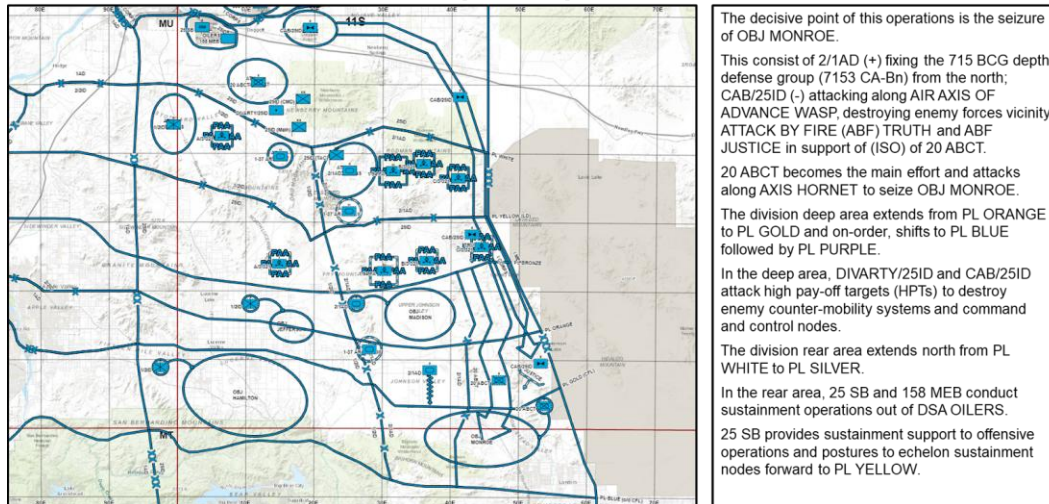


Figure 9. Operation Tropic Tortoise CONOPS.

Enemy Course of Action

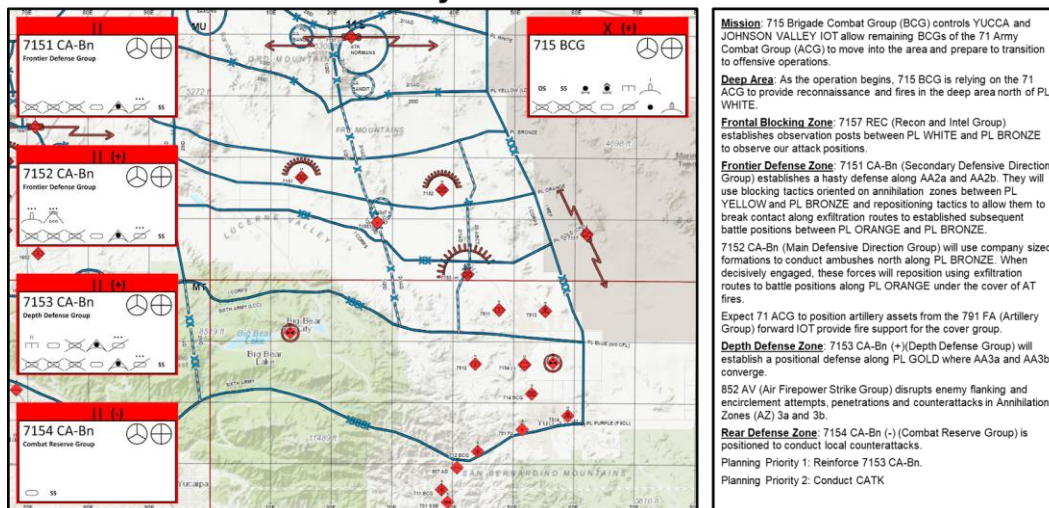


Figure 10. Operation Tropic Tortoise enemy COA.

2. High-Level Findings

The SC3 HE FY25 STP was structured to evaluate the effectiveness of the SC3 ITS, which currently includes the COA-GPT model, the SC3 Sim engine, and the SC3 UI interface, in supporting both individual and collaborative military planning. SMEs participated in a series of structured sessions designed to assess how well the ITS supports mission planning workflows, identifies current limitations, and provides insight into design changes that could produce a more usable and operationally aligned system. This report synthesizes findings across three sessions in relation to the predefined learning directions with particular attention to system usability, output quality, team-based interaction, and performance under realistic planning conditions. Here, we summarize those findings and address design implications and considerations for future research.

Across all sessions, SMEs reported that the current ITS showed meaningful potential, even in its early and partially developed state. SMEs highlighted that completing and connecting these core capabilities should be a high development priority. They noted that, even without the full set of team-oriented collaboration tools and features envisioned for the future, a functional and well-integrated ITS would offer immediate value for individual users. They estimated that the system could reduce the time and cognitive effort required for COA development by approximately 30%, assuming that its core components were completed and effectively integrated.

This 30% improvement on current COA development and mission planning estimate was based on the demonstrated value of COA-GPT's document parsing and seed COA generation providing initial products and plans, SC3 Sim's rapid evaluation of COAs, and SC3 UI's visualization, review, and editing features. SMEs consistently emphasized that starting with a structured baseline COA was one of the most impactful features, as it significantly accelerated early planning and reduced the burden of building from a blank slate. Particularly, the ability to quickly generate a COA, simulate and visualize its results, and revise it through structured editing was expected to reduce manual planning workload and improve the pace and quality of early decision-making. SMEs suggested that this baseline functionality could support more efficient iteration, reduce rework, and help users arrive at viable solutions more quickly. In their view, completing this foundational layer of the ITS would represent a meaningful advancement and a strong platform for building a more collaborative, team-oriented system.

2.1 Summary of ITS Feedback

SMEs expressed moderate confidence in the current version of COA-GPT and strong support for continued development. COA-GPT's strengths were seen in its ability to summarize OPORDs and generate plausible initial COAs. SMEs appreciated the structured prompt interactions and the responsiveness of the model to follow-on edits. However, they also noted limitations in terrain awareness, unreliable handling of user constraints, and output inconsistencies when edits were made to detailed tasking elements. As a result, SMEs reported reluctance to rely on COA-GPT for complex decisions or final COA generation at this stage in development; they indicated that the system would need further evaluation before trusting its outputs in these critical situations where Soldiers' lives could be impacted. They suggested that improving model transparency, justifiability, and tighter integration with simulation to allow users to test and revise model outputs in real planning contexts would help to develop trust in the COA-GPT outputs.

Evaluations of the SC3 Sim and UI were similarly mixed. SMEs noted that the simulation enabled rapid COA visualization and supported the ability to adjust unit tasking and positions, but they also described limitations in temporal granularity, task sequencing, and event traceability. One of the most notable limitations of the simulator was the lack of clarity regarding time in planning; specifically, the simulator lacked clear indications of COA state at each phase of the mission, timelines and locations of key events between phases, and transparency around task execution order. SMEs stated that they could configure COAs but could not easily verify when and how those orders were carried out, whether enabling relationships were respected, or how losses and outcomes were affected. While the simulator was seen as a valuable starting point, the SMEs emphasized that its ability to serve as a meaningful COA testing tool would depend on future improvements to manipulability, explainability, and fidelity of the simulator and its interface.

When evaluated in full team-based planning sessions, the combined ITS demonstrated early signs of value in supporting collaboration, both as a tool for independent taskwork and as a tool for collaborative teamwork. SMEs reported that shared COA generation and plan iteration using the ITS helped reduce manual planning burden and improve team alignment. However, they also highlighted limitations in integration, coordination tracking, and the ability to manage concurrent edits. SMEs indicated that a future ITS should enable distributed teams to manage planning efforts across space and time with clearly defined ownership of planning elements, real-time visibility into changes, and the ability to merge and reconcile edits from multiple COA versions, possibly within or across users, without loss of information. While the current system was not yet perceived as a

reliable teammate, it was described as a promising assistive tool that could be expanded into a full collaborative platform. SMEs suggested that such a system could eventually support more efficient knowledge-sharing, track evolving assumptions, and facilitate consensus-building during time-pressured or high-tempo operations.

SMEs also offered forward-looking guidance on the additional capabilities that would be required to support more complex use cases involving collaborative planning and distributed teams. While the current system shows strong potential for assisting individuals in developing and refining COAs, the long-term vision for the ITS includes enabling teams to build shared understanding, delegate planning responsibilities, and test competing plans across multiple phases and scenarios. The following paragraphs briefly describe the specific features SMEs identified as critical to this goal-state system, including those that would support replanning, branch and sequel plan development, adversarial simulation and wargaming, and collaborative knowledge construction. These capabilities were considered essential for ensuring that the ITS not only improves speed and usability but also strengthens the collective reasoning and adaptability of military planning teams.

SMEs postulated that a fully realized, goal-state ITS could reduce planning time and effort by a minimum of 60% to 70%, provided it was developed with the features and workflows necessary to support real-world decision-making. Among the most important features identified were the abilities to plan, simulate, and visualize COAs within discrete temporal phases. Although the COA-GPT component of the ITS had preliminary functionality to plan in phases, the integration with the simulator lacked support for inspection of discrete phases, which was a notable drawback. SMEs indicated that phase-based planning and assessment would be essential for coordinating unit actions within time and space, to support reasoning across cause-and-effect timelines, and for structuring COA comparison across multiple mission phases. They emphasized that users should be able to build and adjust COAs on a per-phase basis, simulate up to key decision points, and inspect outcomes tied to temporal triggers and contingencies. SMEs also strongly recommended the ability to wargame against a human adversary. They described this as useful for testing the plausibility and adaptability of plans, and for encouraging more realistic and rigorous COA development. A functional wargaming capability was described as one in which each user could control opposing forces; adjust COAs through branch and sequel plans from decision points or replanning actions; and have the abilities to freeze, replay, walkthrough, or annotate outcomes during simulation. This, SMEs explained, would allow users to explore the dynamics of maneuver, deception, and counter-decision-making in ways that static COA comparisons do not currently support.

For supporting team planning, SMEs identified structured delegation of planning authority as an important feature for supporting collaborative and distributed use. They suggested that the ITS should allow users to assume ownership over specific organizational levels (e.g., battalion vs. company), geographic domains (e.g., distinct sectors of the AO), and tasking roles (e.g., fire support, maneuver, logistics). These role configurations, they said, must support both synchronous and asynchronous workflows and reflect the natural division of labor found in planning staffs. SMEs emphasized that a successful system would maintain persistent visibility into which user controls what, when changes are made, and how those changes affect downstream dependencies. Several other features were also identified as important for a successful, goal-state ITS, which were focused more on the user experience of engaging with the technologies. SMEs suggested that the natural language interaction with COA-GPT should be expanded to better support iterative refinement of specific COA elements. This included the ability to modify tasks for a specific unit in a specific phase using natural language instructions such as “delay this movement until after the breach is completed” or “remove the recon element from this axis and replace it with a support-by-fire position.” SMEs also stated that the system should enable direct plan manipulation by unit, objective, and mission phase. Full traceability from each COA recommendation to the source document or rationale was another frequently mentioned need. SMEs described how this would support auditing, explanation, and doctrinal validation. They also noted that users would benefit from tools that allow rapid switching between tactical and operational perspectives, drag-and-drop unit organization and tasking, live feedback of measures of interest during editing, and visual mapping of task interdependencies across units and phases in time and space.

The vision for a goal-state ITS was not limited to automation or visualization enhancements. SMEs described a system that would fundamentally reshape how teams interact with digital planning tools. In their view, a successful ITS would allow multiple users to collaboratively edit a plan in real time and asynchronously, remotely and collocated, with full transparency into roles and changes. It would support persistent COA comparisons, live task updates, and synchronization across planning elements. It would also need to support a shared structure of ownership over planning elements (such as unit control, spatial areas, and functional tasking) in a way that reflects how command post operations are actually run. SMEs emphasized that this structure is essential for supporting distributed, multirole planning and reducing the coordination overhead that otherwise limits tempo.

2.2 Future Research

Looking ahead, a key objective for the SC3 research effort is to determine how best to evaluate the ITS in team-based planning environments, both in terms of effectiveness and accuracy of the system as well as in relation to current MDMP processes. Although much of the touchpoint focused on understanding the system's future design requirements and user needs, the next phase of work must also focus on how to structure experimentation that produces meaningful, transferable insights. This includes not only testing ITS performance in relevant operational contexts, but also validating our data collection strategies, team structures, scenario designs, and C2 processes. A well-structured research approach will ensure that future findings are rigorous, comparable, and relevant to the real-world planning environments the ITS is intended to support. This will also be critical for studying and informing future requirements across all facets of capability development: Doctrine, Organization, Training, Materiel, Leadership and Education, Personnel, Facilities, and Policy. One of the strongest takeaways from SME feedback was the importance of testing the ITS in realistic team settings. SMEs repeatedly emphasized that COA development does not occur in isolation, but is carried out by teams with assigned roles, distributed responsibilities, and shared mental models. To meaningfully evaluate the impact of the ITS, experiments should involve fully "staffed" planning teams comprising either military personnel or trainees who are sufficiently familiar with, and capable of, conducting MDMP. These teams should be organized to reflect operational roles (command, logistics, maneuver, intelligence, support, etc.) but must also be adaptive to the "art of the possible" for staff organization that these novel technologies will enable. Whether drawn from training pipelines or operational units, the teams must be capable of producing realistic planning behavior and interdependent workflows that mirror actual mission planning practices and user skillsets.

A critical component of future research on the SC3 ITS will be the establishment of valid task and evaluation baselines. Teams should begin by developing COAs for at least two (ideally more) distinct operations using traditional planning methods. These baseline missions should differ meaningfully in planning focus such that they require teams to employ components of the ITS to varying degrees. For example, one may emphasize terrain and combat coordination, while another may focus on logistics, timing, and resource management. What will matter most is that the missions represent diverse planning demands and that the teams perform them using the same procedures currently taught or expected in operational environments. Baseline data should capture not only planning outcomes but also team processes, communications, sequencing and timelines, indicators of shared understanding, behavioral indicators of coordination, etc. This will allow future

studies to directly assess the value added by the ITS in comparison with how teams currently work without it. The scenarios should maintain comparable structure and planning demands but must be distinct enough that teams cannot simply replicate their previous COAs to address the scenario. This will allow for meaningful comparisons of planning outcomes and decision-making behaviors, while also validating some degree of generalizability of findings to novel scenarios. With two mission types and two planning approaches (traditional and ITS-supported) this foundational experimental design enables controlled comparison of how the ITS influences planning across different types of operational challenges. This approach will also allow researchers to identify which elements of the ITS are most beneficial for different types of planning, such as unit task coordination or timeline optimization.

To build on the traditional-versus-ITS focused foundation, which aligns more with a planning-centric future operations cell, experimentation should be extended to include scenarios with midcourse changes in mission requirements or information, prompting the need to reconsider the plan, which conversely aligns more with the execution-centric current operations cell. These perturbations could include revised commander's intent, altered unit operating conditions, new intelligence, or changes in terrain. The goal of these scenarios is to evaluate how teams adapt, replan, and manage cognitive load under pressure. Crucially, these additional scenarios should still follow from the baseline ITS experimental design. Organizing experimentation in this way ensures that data remains compatible across studies, supports longitudinal learning, and avoids the trap of disconnected, one-off demonstrations that are difficult to generalize. Scenarios can be structured to reflect operationally plausible disruptions but should be integrated into a consistent experimental framework.

Finally, successful experimentation depends on participants being prepared to use the ITS effectively. Research protocols should incorporate clear role assignment and structured training tailored to each user's responsibilities within the planning team. Each participant must be trained not only on how to use the interface, but how to use it in the context of their assigned role. Without sufficient training, it becomes difficult to separate the impact of the ITS from the challenges of learning the system itself. To ensure the integrity of the research, training should be designed to prepare each user to operate as a fully functional member of a planning team during ITS-supported scenarios, with a progressive curriculum to allow for sufficient training within constraints of a limited experimentation timeline.

In sum, the touchpoint demonstrated that the ITS is on a promising trajectory, with meaningful capability already in place and a clear pathway toward impactful operational use. SMEs estimated that, once the currently demonstrated features are

fully completed and integrated (such as closed loop iteration between generative AI and constructive simulation, collaborative COA editing, and granular spatiotemporal COA review and refinement) the ITS could reduce the time and cognitive effort required for COA development by approximately 30%. They emphasized that even without advanced team features, this foundational level of functionality would provide significant value for individual users, particularly in fast-paced or resource-limited environments. SMEs described the system's ability to generate an initial COA and support rapid, structured and data-driven iteration as a major advancement over current manual processes. Completing the development of these core capabilities should likely remain the highest priority. Doing so will not only provide immediate planning benefits but also solidify the essential platform on which more complex, team-oriented capabilities can be built.

Looking forward, SMEs outlined a clear and compelling vision for a goal-state ITS that could support collaborative planning at scale. They suggested that a fully realized version of the system could reduce planning costs by 60% to 70%, provided it includes iteration and team focused features such as phase-based planning, adversarial wargaming, role-based tasking, and transparent replanning. Importantly, they also emphasized that evaluating the ITS in these contexts will require rigorous experimentation, grounded in realistic team-based planning conditions. Establishing valid baselines, training teams in role-specific ITS use, and introducing controlled scenario variations (e.g., changing commander's intent or resource constraints) will be necessary to determine how the ITS affects not only planning outcomes but also team cognition, adaptation, and decision-making. With the right balance of focused development and structured research, the ITS has the potential to transform not just how quickly plans are built, but how effectively teams reason, communicate, and respond under pressure.

2.3 Most and Least Valuable Features of the ITS

Across all sessions, SMEs expressed support for the direction of the ITS research and noted that the system is steadily improving. While they agreed the tool is not yet ready to support live COA development, they emphasized that its current capabilities already offer value, especially for individual planners working without a full staff. The system helped users move more quickly through early planning steps and provided a structured starting point for COA development. SMEs saw clear potential in the integration between COA-GPT and the SC3 Sim. Even in its current state, the system enabled users to move from mission analysis to COA sketching and simulation, helping to frame planning decisions. However, several participants encountered mismatches between what COA-GPT said it changed and what appeared in the simulation. These disconnects limited trust in the system and

highlighted the need for more consistent and transparent synchronization between planning edits and simulation updates.

The most critical missing feature identified was the lack of support for phasing and task sequencing. SMEs repeatedly noted that without the ability to define actions over time and across operational phases, it becomes difficult to validate whether a COA is executable. They recommended adding timeline-based tasking, phase-aware visualizations, and simulation playback that allows planners to see how operations unfold step by step. Participants also suggested additions that would support more realistic wargaming. These included incorporating both friendly and enemy COAs, defining evaluation criteria, and enabling comparisons across multiple plan options. SMEs asked for COA-GPT to be expanded so it can propose alternate task organizations or recommend the creation of task forces when appropriate, based on higher-level mission demands.

Looking ahead, SMEs encouraged future touchpoints to include full scenario setups with real-world structures, realistic operational constraints, and clear opportunities to test how the ITS supports planning under pressure. One participant summarized the general feeling well: the system is not yet where it needs to be, but it has come a long way and continues to move in the right direction. With continued focus on sequencing, coordination, and output clarity, the ITS has the potential to make planning faster, more consistent, and easier to adapt.

2.3.1 What ITS Features Were Useful and How?

SME feedback across both sessions highlighted several ITS features as immediately useful and aligned with operational needs. The most positively received components fell into three major categories: 1) ease of use and intuitive interface, 2) simulation and tasking features, and 3) COA-GPT's document parsing and summarization capabilities.

2.3.1.1 Intuitive and Accessible UI

Across all sessions, SMEs consistently emphasized that the ITS was easy to learn and use, even in its partially developed state. Participants (Alpha and Golf) specifically noted that the point-and-click interface for directing unit movement was straightforward and efficient, especially when grounded in a terrain-aware map. One SME noted the value of having known terrain features visible, which helped orient planning decisions and simulate real-world operational reasoning. The interface also enabled quick task editing and reassignment using direct interactions rather than complex menus. Alpha described the ease of adding new tasks and moving units, while Echo praised the interface's capacity to visualize COA execution paths and provide immediate status feedback after simulation runs.

SMEs also saw potential in the augmented map concept, particularly for enabling shared understanding across distributed command nodes, signaling that the ITS may offer unique advantages in geographically dispersed planning contexts.

2.3.1.2 Constructive Simulation and Plan Execution Feedback

The simulation component was described as a major strength, particularly for its reusability and responsiveness to changes. Golf highlighted the benefit of being able to repeatedly run the simulation with just minor edits, accelerating iterative COA refinement. Alpha and Echo also highlighted the simulation's ability to quickly reflect the outcomes of input COAs and automatically display results such as friendly and enemy losses, which enhanced user understanding of potential operational tradeoffs. In particular, the pre-seeded simulation was valued by participants as a mechanism to avoid manual setup and "drag-and-drop fatigue." Echo appreciated that the baseline scenario already included starting conditions, allowing planners to immediately begin evaluating and modifying courses of action. This strongly supports the hypothesis that automation of scenario initialization is a key value-add of ITS technologies.

2.3.1.3 COA-GPT's Summarization and Planning Support

SMEs provided positive commentary on COA-GPT's ability to ingest large amounts of operational data and surface key information quickly. Alpha called this out explicitly, identifying it as valuable for speeding up information processing during the planning phase. Golf also noted that the model continues to improve and already provides plausible and modifiable COA suggestions. This function appears particularly useful for accelerating the initial stages of COA development, providing users with a viable starting point rather than a blank slate.

The integration of COA-GPT with structured tasking workflows was seen as foundational to enabling rapid iteration, and several participants (e.g., Echo) suggested that its outputs made the follow-on simulation more actionable and effective.

2.3.2 What Are the Three Most Significant Strengths of COA-GPT Identified by Users?

2.3.2.1 Speed and Efficiency in Parsing and Synthesizing Operational Inputs

Participants praised COA-GPT's ability to quickly absorb and summarize large volumes of mission-related data, including commander's intent, mission tasks, and synchronization requirements. This functionality was highlighted as particularly valuable for accelerating the early stages of COA development. One SME noted

that COA-GPT's output offered "a good start" and saved considerable time compared with starting from a blank slate. The system was described as helpful in transforming mission analysis into an initial structured plan with minimal user prompting, enabling individual planners to move more quickly into task organization and refinement.

2.3.2.2 Natural Language-Based Plan Modification

The ability to use natural language instructions to iteratively refine COAs was another standout strength. In session 3, participants interacted with COA-GPT to make phase-specific edits to the seed COAs, such as repositioning units or adjusting their tasking sequences. Although the system did not always execute instructions flawlessly, SMEs expressed that the capacity to prompt changes like "move this unit after the breach" or "add artillery prep before maneuver" represented a promising and intuitive interaction mode. This kind of interface allowed them to engage in the planning process in a manner that mirrored natural staff communication, reducing the cognitive burden of manipulating plans through menus or abstract inputs.

2.3.2.3 Support for Generating Structured, Phase-Based Plans

SMEs recognized COA-GPT's emerging capability to structure plans by temporal phases and associate unit actions with specific points in the operation timeline. While this functionality was still under development and not always fully synchronized with the simulator, users appreciated that COA-GPT could at least begin to build phased plans, distinguishing between early shaping operations and main efforts. This structuring was seen as critical for operational realism and mission execution as it helps planners track dependencies, establish decision points, and ensure doctrinal alignment across phases.

2.3.3 What ITS Features Were Problematic and How?

SMEs were generally positive about the ITS and its potential to support planning workflows, but they also identified several recurring usability limitations and capability gaps. Feedback focused on three major problem areas: 1) limitations in COA-GPT accuracy and planning logic, 2) challenges with simulation structure and COA comparison, and 3) friction in the UI for planning coordination and clarity.

2.3.3.1 COA-GPT Limitations in Terrain Awareness and Planning Validity

Across both sessions, SMEs noted that COA-GPT did not adequately account for terrain constraints or doctrinal priorities in its generated COAs. Alpha highlighted that its current inability to incorporate terrain meaningfully limits its usefulness in

realistic COA generation. Golf stated that the COAs lacked viability at the COL level and missed key operational objectives: for instance, omitting OBJ MONROE, which was identified as the division's main effort. This undermined SME trust in the tool and its potential for producing plausible plans under realistic conditions.

2.3.3.2 Simulation Structure and COA Comparison Limitations

The constructive simulation was praised for its potential, but current implementation lacked several features critical for operational planning. Alpha noted the inability to coordinate actions across phases or visualize operations over time, which made it difficult to assess temporal synchronization. Both Echo and Golf flagged the lack of persistent COA comparison tools as a major issue. Users could not easily save, reference, or compare COAs within the interface, requiring external note-taking and making iterative refinement cumbersome. Echo also noted that COAs did not always satisfy critical directives from the OPORD, compounding trust concerns.

2.3.3.3 UI Complexity and Control Friction

Numerous UI-level challenges were raised across both sessions. SMEs noted that task insertion and unit-task alignment were often confusing or restrictive (Alpha), with issues like the system resetting task locations to the unit position rather than a prior task. Echo and Golf emphasized that the system became harder to navigate as the number of units increased and the lack of aggregation mechanisms (e.g., moving a battalion as a whole) created unnecessary friction for formation-level planning. Display clutter was another concern: SMEs asked for ways to toggle off visual elements like movement paths and range rings that interfered with focus. Additionally, default UI behaviors (such as all layers reappearing after navigating away from a view) created repetitive cleanup work and reduced efficiency (Golf).

Across the board, SMEs conveyed that many of these issues did not reflect a fundamental design flaw but rather areas in need of refinement, prioritization, and increased configurability. The presence of these friction points did not negate the system's value, but did constrain its current effectiveness, especially for more complex planning tasks or dense operational scenarios.

2.3.4 Which Element of the ITS is Most Lacking, Specifically in Terms of Hindering Your Ability to Conduct Planning in a Sufficiently Realistic Manner?

The most significant limitation of the ITS, as reported by SMEs, was the lack of support for temporal sequencing and phase-based coordination. Without the ability to break operations into time-based phases or define task dependencies, the system

prevents realistic modeling of how units interact over time. SMEs explained that this limitation undermines the doctrinal validity of the planning process. For instance, in one scenario, a brigade was shown seizing a key objective before supporting artillery units had moved into position, violating standard sequencing. This made it difficult for users to determine whether COAs were actually executable. SMEs recommended integrating a timeline-based planning structure that allows tasks to be assigned by phase, ordered by dependency, and visualized with clear operational timelines. They also requested simulation playback tools aligned with these phases, so that planners can evaluate branch points and decision triggers in context.

A second major issue was COA-GPT's inability to account for terrain, operational priorities, or unit constraints in a consistent and realistic way. Several SMEs noted that the generated COAs ignored critical objectives or proposed implausible unit taskings. For example, one COA failed to mention OBJ MONROE, the division's main effort, casting doubt on the utility of the output. While SMEs appreciated COA-GPT's ability to summarize orders and generate structure, they emphasized the need for better grounding in mission inputs and more transparent justification for its recommendations. They advised incorporating checks for doctrinal completeness, ensuring that all specified and implied tasks are addressed, and improving integration with terrain overlays and order references.

The UI also presented some challenges, especially at scale. SMEs described cluttered displays, difficulty tracking unit status, and inefficient controls when working with large formations. Some expressed frustration at having to move units individually rather than as a combined element and not being able to filter out irrelevant visual elements, like routes from nonessential units. Suggested improvements included tools for grouping units by command structure, toggling overlays, searching for specific units, and quickly switching between COAs or planning perspectives.

2.3.5 What Are the Three Most Significant Areas of COA-GPT for Improvement Identified by Users?

2.3.5.1 Poor Integration of Terrain and Doctrinal Constraints

Across both qualitative comments and numerical ratings, SMEs consistently cited COA-GPT's failure to meaningfully incorporate terrain features and doctrinal planning principles. Users reported that COAs often ignored key spatial boundaries like phase lines, omitted preconditions for maneuver actions, and failed to align sequencing with standard combined arms practices. For instance, one SME noted that COA-GPT recommended movement toward OBJ MONROE without first

achieving preconditions at OBJ MADISON, which is an error that contradicts operational logic. These issues eroded trust in COA-GPT's ability to generate executable, mission-valid plans and suggest a need for tighter integration of terrain models, doctrinal rules, and constraint checks into the generation process.

2.3.5.2 Inconsistent Responsiveness to Natural Language Edits

Although natural language interaction is one of COA-GPT's standout positive features, its current implementation was also viewed as somewhat unreliable and unpredictable. Participants in session 3 attempted to use natural language to adjust unit assignments, change task timing, and add phases, but sometimes found that the model either failed to implement the change or reported success without making visible adjustments. For example, Echo reported that COA-GPT deleted all phases except Phase 1 after a prompt and failed to fully reimplement the missing phases when prompted again. Another user (Alpha) noted that while edits to 1st Squadron, 1st Cavalry Regiment (1-1CAV) and 5th Battalion, 20th Infantry Regiment (5-20IN) were initially successful, subsequent changes to 1st Battalion, 6th Infantry Regiment (1-6IN) not only failed but disrupted unrelated units' plans. These issues suggest the need for better prompt parsing, validation, and confirmation feedback loops such as displaying a summary of changes made or requesting clarification when ambiguity is detected.

2.3.5.3 Low Reliability in Task Sequencing and Synchronization

Participants found that COA-GPT repeated tasks across some phases, skipped critical handoffs, or misaligned task execution with mission timelines. Alpha explicitly noted that some tasks were repeated across phases when they should have transitioned, and Echo described cases where task assignments were incompletely or incorrectly reflected in the synchronization matrix. This reduced user confidence in COA-GPT's ability to maintain logical progression and interunit coordination across phases. These deficiencies were reflected in low average scores on items assessing time-based phase management (2.0/6) and alignment with mission requirements (2.66/6), pointing to a systemic issue in temporal reasoning and task handover representation.

2.4 ITS Priority Feature Gaps

Across all evaluation sessions, SMEs consistently highlighted a clear vision for what a fully realized ITS should support: rapid, structured, and doctrinally valid planning—both individually and collaboratively. The highest-rated design features in the evaluation reflect this vision, with particular emphasis placed on temporal control, intuitive tasking workflows, adversarial simulation, and collaborative role

management. Importantly, no queried design feature received an average rating below 70/100, underscoring the perceived value of each proposed improvement (Figure 11). Quantitative data was not captured for Alpha due to an error, so their ratings are omitted from the results presented in Figure 11.

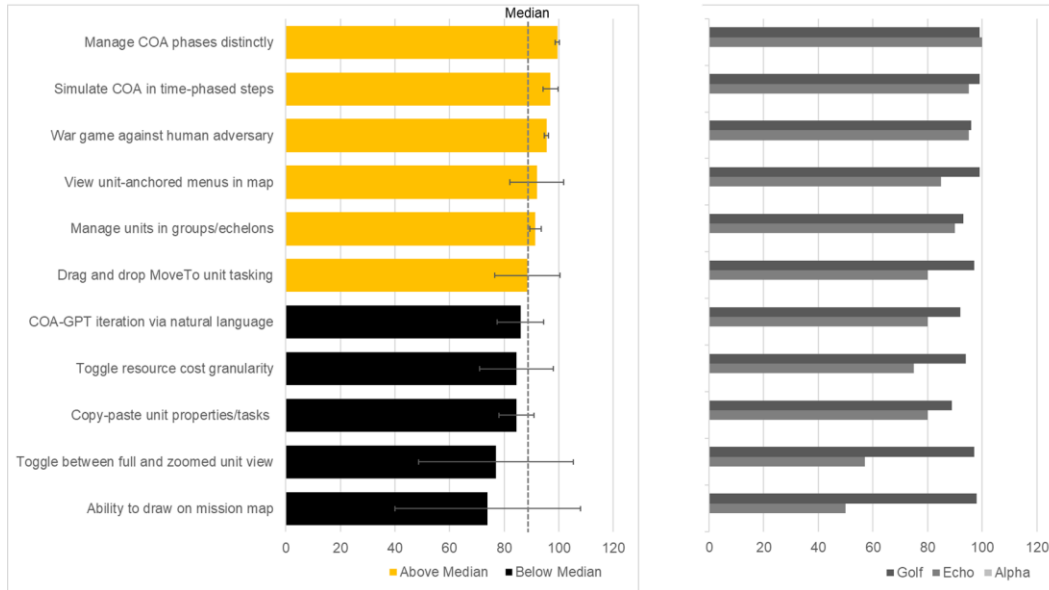


Figure 11. Session 2 design feature ratings.

2.4.1 Temporal Structuring and Phase-Based Simulation

The most strongly emphasized feature gaps were those related to temporal elements of the COA, particularly the ability to modify and display the timing of events within the UI with greater granularity and interpretability. The ability to manage COAs by phase (Figure 11, Item 1) or to simulate operations up to key decision points or by phase (Figure 11, Item 2) received the highest ratings. SMEs described how essential temporal sequencing is for modeling cause-and-effect chains in operations such as ensuring artillery has cleared a position before maneuver forces advance. They proposed functionality that enables planners to define tasks and dependencies by phase, simulate plans incrementally, and observe the timing of critical failures (e.g., when supporting units are destroyed or cannot execute follow-on tasks). One participant noted that a system lacking this ability leads to COAs that violate basic doctrinal logic, such as seizing an objective before supporting fires have been emplaced.

To fully support these use cases, SMEs recommended the following:

- Adjusting task assignment timing of multiple units with greater ease of use and precision.
- Toggling and visualizing simulation playback by phase or more granularly.

- Marking decision points and contingency triggers.
- Creating and naming unit-level control measures like echelon-specific phase lines, no fire zones, and limits of advance, in addition to those inherited from higher echelons.

2.4.2 Streamlined Tasking and Unit Control

SMEs also emphasized gaps in manipulation of tasking and units, specifically the ability to group and manage units when assigning tasks at different echelons, rating it high (Figure 11, Item 5), which reflects frustration with the current system’s requirement to manipulate units individually. SMEs expressed a strong preference for functionality that allows moving battalions or other formations as cohesive groups, aligning with real-world command structures.

Closely linked, and also highly rated, was the ability to edit unit properties and tasking from within the map view (Figure 11, Item 4). SMEs suggested a concept for clicking directly on units, rather than navigating dropdowns or nested menus, would allow faster, more intuitive task assignment. Drag-and-drop tasking, with timing constraints (Figure 11, Item 6), was also rated highly and seen as important for rapid plan creation, particularly when coordinating movement tasks and synchronizing unit arrivals with mission phases.

Additional features in this category included the following:

- Copy–paste tasking across units to reduce repetitive actions for similarly assigned units.
- Zoom toggling between unit-level and operational views to support contextual awareness while editing.
- Toggleable resource cost visualization for sustainment and resourcing insights.

SMEs further recommended integrating search and follow functions (e.g., “Where is B-1-1?” or “Track D-47 during Phase 2”), highlighting how situational awareness suffers when users must scroll or pan manually to track unit behavior.

2.4.3 Natural Language Editing and COA-GPT Refinement

SMEs rated highly the need to expand natural language editing via COA-GPT (Figure 11, Item 7), using plain language to make specific, targeted changes to COAs, such as “Delay movement of C 1-1 until Phase 3.” This interaction mode was seen as essential for enabling users to refine COAs at speed without manually rebuilding task sets in the interface.

However, SMEs also noted limitations in COA-GPT's current responsiveness, especially when attempting edits tied to specific phases or geographies. They recommended expanding traceability, such that users can click on a unit or task and ask, "Why did you do this?," providing clear links from each generated action back to the source order or doctrinal justification. This would greatly improve transparency and support doctrinal validation during planning.

2.4.4 Adversarial and Collaborative Planning Support

The ability to wargame against another human adversary, rather than treating the simulation as a static validator of blue plans, was also rated highly (Figure 11, Item 3). SMEs stated this would be critical for improving realism, adaptability, and rigor of COA testing, and they envisioned a dynamic planning environment where users could do the following:

- Control opposing sides of the battlefield.
- Introduce deception, maneuvers, and counters.
- Pause, annotate, and revise COAs during execution.
- Replan in response to enemy decisions.

These features would directly support training and decision rehearsal use cases and better reflect the adversarial nature of real-world operations.

For collaborative planning, SMEs emphasized the importance of structured role and area ownership. They proposed that users should be able to assume responsibility for geographic areas (e.g., north sector of the AO), organizational elements (e.g., battalion task force), or functional domains (e.g., fires or logistics). This would enable distributed planning, mirror command post organization, and allow both synchronous and asynchronous workflows. SMEs stressed that all changes must be traceable by user, with clear visibility into who made what edits and when, to ensure accountability and coordination.

2.4.5 Additional SME Recommendations

In their open-ended comments, provided during planning sessions and AARs, SMEs suggested several additional enhancements to the ITS:

- A "confirm task org" prompt or checklist to validate unit roles and tasking completeness.
- Notification when the simulation completes running.
- Naming of all graphics and control measures.
- Support for engineering and sustainment functions, such as obstacle placement, resupply, and casualty replacement.
- Search tools for quickly locating units or assessing task status.

- Visualization of when units failed or succeeded in executing their missions, aligned to time and space.

Collectively, this feedback indicates a common theme that must be central to future development of these systems: fluid, well-designed UIs and intelligent, transparent decision support features that aid in identifying and presenting the effects of planner decisions are just as critical to the success of these technologies as the capabilities to rapidly generate or rollout COAs. If these technologies are to provide any benefit to planning, they cannot be relegated to secondary or tertiary considerations. No matter how rapidly produced or analyzed, COAs that are not easy for planners to understand and modify will not support them in future C2 processes.

2.4.6 Priority Simulation Mechanics Design Goals

In addition to the above AAR SME feedback on the current prototype implementation of the system, the research team also identified several priority areas for future development, focused primarily on updates to the simulation mechanics, that were the topic of the collaborative design discussions during the touchpoint. These areas of future improvement are as follows:

- Representation of units in terms of constituent classes with more granular tracking of their properties and state.
- Representations of multiple armament capabilities and maximum effective range within units.
- Attrition rates that vary as a function of proximity between engaged units.
- Resource consumption and simplified logistics and resupply mechanics.
- Indirect fire mechanics that are dependent on observation by separate units.
- Terrain occlusion of observation and combat.
- Terrain classification into many types with varied and unit-specific impacts on navigation, mobility, and observability.

More detailed descriptions of how we will approach these simulation mechanic improvements are provided in Section 3.3.6. Once implemented, these improvements to the constructive simulation mechanics of SC3 Sim will yield a richer and more accurate prediction of COA outcomes. The aim of each of these improvements is to provide a forecasting tool that better aids planners in their analysis and refinement of COAs, particularly supporting planning for additional warfighting functions represented within planning staff and MDMP.

However, with each of these potential features, implementation will need to be considered and balanced against computational cost and its effect on simulation speed, as this is a fundamental element of SC3 Sim. Furthermore, the ability to toggle on or off the more computationally costly mechanics should be pursued

(where possible, from a software architecture perspective) so that a simulation fidelity can be dynamically adjusted during the COA design process, based on tradeoffs in exploration and exploitation of COAs.

In addition to these more near-term capability improvements, further-term capabilities we aim to pursue include the following:

- Horizontal scalability to enable parallel rollout of many simultaneous simulation instances.
- Seedable, pseudorandom stochasticity to produce statistical distributions of outcomes.
- User extensibility to better define unit and tactical mission task configurations.
- Nested behavior tree for more granular definition of unit actions.
- Defined terminal conditions to capture task and purpose.
- Partial observability for fog of war for both friendly and enemy forces.
- World models for each unit to represent denied, disrupted, intermittent, and limited communication environments.

These advanced simulation capabilities are critical to fully realizing our vision for the ITS and the SC3 program as a whole. They will enable a planner and AI to rapidly analyze and refine COAs through a rigorous and adaptable simulation platform.

2.5 ITS Usage Characteristics

During the touchpoint, participants engaged with the ITS in a hands-on manner that revealed both its utility and current limitations for supporting mission planning tasks. Although the evaluation did not simulate a full staff environment, the system enabled individual planners to quickly step through the early stages of the military decision-making process. Participants appreciated having pre-generated COA options from COA-GPT, which provided a structured starting point rather than requiring them to build a plan from scratch. This initial boost was particularly valuable for streamlining early COA development and helped users orient themselves to the mission without extensive document parsing.

Most interaction periods followed a consistent pattern. Participants selected one of the seed COAs, ran the simulation to observe outcomes, and then iteratively refined the plan using the edit positions interface. Plan edits were often driven by a single high-level metric: total friendly versus enemy unit losses. This encouraged a “test and tweak” workflow, where users manually adjusted unit positions and tasking, reran the simulation, and evaluated whether the changes reduced overall friendly

casualties. While this cycle provided a functional mechanism for iterating on COAs, participants also acknowledged that it simplified the assessment of COA quality. In the absence of tools to view outcomes by phase or clearly trace unit losses back to specific actions or conditions, users were often forced to rely on trial-and-error adjustments. Some even described their process as “gaming the system,” focused more on manipulating loss counts than improving doctrinal alignment or tactical soundness.

Accordingly, the majority of user time was spent in the edit positions panel, where units could be retasked, repositioned, or re-sequenced. The COA analysis panel was used primarily to check summary outcome statistics between simulation runs. Participants noted that certain elements (e.g., defensive postures, support sequencing, or task dependencies) were occasionally difficult to interpret within the simulation, which limited their ability to reason about the cause-and-effect dynamics of their plans. Despite these limitations, the ITS supported a meaningful and productive interaction pattern. It allowed users to engage directly with planning artifacts and observe the consequences of their decisions, laying a strong foundation for further refinement. With future improvements to phase-based playback, outcome traceability, and doctrinal visualization, these interactions could become even more informative and aligned with real-world planning practices.

We quantitatively assessed planning effectiveness, adaptability, and engagement characteristics with SC3 Sim and UI, analyzing COA design quality (simulation outcomes for combat power, losses, etc.), interactions patterns (UI page transitions, mouse pointer interactions, etc.), and attentional indicators (gaze trajectories, pupil dynamics, etc.) across the three users (Alpha, Echo, and Golf). At a high level, while the overall pattern of usage described above held across the users, this analysis suggests distinct user strategies and behaviors, as follow.

Regarding COA design quality, Alpha demonstrated more aggressive COA design tactics, with sharp shifts in force outcomes often occurring later in the planning sessions. Echo showed stable patterns with intermittent engagement spikes, suggesting a flexible but controlled planning style. Golf consistently achieved favorable Blue/Red ratios with minimal friendly losses, indicating a methodical and effective approach.

Regarding interactions patterns, Alpha demonstrated a methodical transition pattern between AOIs, balancing time between COA Creation and Edit Positions with growing use of analysis tools over time. They showed the broadest and most variable pointer behavior, with widespread cursor dispersion and fluctuating straightness across AOIs, indicative of high interaction effort and iterative planning strategies. Echo engaged broadly across all AOIs with frequent transitions,

reflecting a flexible and exploratory planning strategy. They exhibited moderate pointer efficiency with frequent shifts across AOIs, reflecting a flexible and adaptive approach, with patterns revealing exploratory activity, particularly in the COA Analysis AOI. Golf remained largely in the Edit Positions AOI, indicating a focused, execution-oriented approach with minimal interface exploration. They consistently demonstrated high pointer straightness and tightly clustered cursor paths, especially in the COA Creation and Edit Positions AOIs, suggesting a structured and efficient interaction style with minimal exploration.

Regarding attentional indicators, Alpha demonstrated the most visually active profile, with frequent, rapid shifts in gaze (high saccade count and velocity variability) and many brief fixations, suggesting a broad visual scanning strategy and fast-paced attention shifts across interface elements. Alpha's larger pupil size and low blink rate are markers associated with elevated cognitive workload and continuous attentional engagement. Altogether, the eye-tracking profile of Alpha reflects high visual vigilance and active monitoring behavior. Echo demonstrated a moderately active visual profile, based on moderate eye saccade count, velocity variability, and fixations, indicating a balanced pattern of gaze shifts and more sustained attention on individual interface elements. Their pupil size and blink rate were markers of reduced or fluctuating cognitive workload. These patterns suggest a measured but less sustained engagement strategy, with periodic attentional breaks that may reflect a more flexible or reactive visual monitoring approach. Golf showed the least visually active profile, with minimal gaze shifts, narrow visual field, and few and lengthy fixations, suggesting a slower attentional rhythm with sustained focus on fewer interface elements. Small pupil diameter and elevated blink rate were markers of lower cognitive workload and reduced engagement. The eye-tracking profile of Golf suggests a more passive and constrained visual approach, potentially reflecting reduced task demand or a more simplified interaction strategy.

2.5.1 What Impact Do You Think ITS Had on Your Planning Process?

Although the ITS was not evaluated under conditions equivalent to a full staff COA development process, participants indicated that it still offered meaningful support for individual planners during early phases of mission planning. The combined use of COA-GPT and the SC3 Sim helped users move through key planning steps more quickly than they otherwise might on their own, especially by providing a structured starting point for COA development. One SME noted that they were able to reach multiple MDMP phases more rapidly with ITS support, even though they were working without the benefit of a full planning team.

The ability to generate and test an initial COA sketch through the linked modules of COA-GPT and the simulator was seen as a valuable foundation. SMEs appreciated that COA-GPT could synthesize mission inputs and propose an initial COA and the simulator could provide a basic operational run-through of the plan. However, several limitations emerged in how useful and trustworthy those outputs were for deeper planning. For example, one SME described situations where COA-GPT claimed to have completed a task that, on review, was not accurately reflected in the plan. Others encountered challenges with editing COAs and ensuring that updates were faithfully implemented within the system. These issues reduced confidence in the outputs, especially in terms of how well they would hold up under scrutiny in a collaborative or operational setting.

Despite these concerns, the ITS was widely described as a promising and rapidly maturing tool. Participants agreed that, for individual planners working in time-constrained or resource-limited contexts, the system already offers a valuable head start and a clear structure for further refinement. One SME summarized this sentiment by noting that while the ITS is not yet capable of replacing full traditional staff processes, it is already a strong starting point and has made significant progress. With continued development focused on improving output accuracy, task traceability, and editing responsiveness, the ITS has the potential to meaningfully accelerate and support COA planning both for individuals and, eventually, for full teams.

2.5.2 How Satisfied Are Users with COA-GPTs Performance in Facilitating MDMP?

COA-GPT was broadly viewed by participants as a valuable asset for accelerating the early phases of the MDMP, especially in its ability to generate an initial COA from complex input data. One SME described it as offering the benefit of “starting with something on the page,” which significantly reduced the time and effort typically required to parse operational orders and convert them into actionable plans. Rather than beginning with a blank slate, users were able to quickly move into reviewing and refining a base plan, which provided momentum to the planning process. Another major benefit of COA-GPT was its capacity to digest dense planning inputs and present a coherent output that aligned, at least in structure, with expected operational phases and doctrinal components. Participants noted that the model could identify key mission elements and propose plausible unit assignments and synchronization. This made it easier for users to conduct initial planning and orientation activities, particularly without a full staff. The ability to use natural language for prompting and revision was also seen as promising, supporting a more intuitive and accessible form of interaction.

Despite these strengths, current system limitations prevent COA-GPT from fully realizing its potential as an MDMP accelerator. Users reported multiple instances where the tool failed to make intended edits, misunderstood the scope of requested changes, or misrepresented those changes in the simulation. These issues affected user confidence, especially when outputs required manual inspection or reimplementation to ensure accuracy.

Although COA-GPT is not yet ready to replace traditional planning tools, it is already supporting faster early-stage COA development and improved user engagement. Participants agreed that with improvements to edit handling, constraint awareness, and traceability, COA-GPT could become substantially easier to use and more effective than current systems. With continued development, it has the potential to reduce planning time and help teams produce higher quality COAs.

2.5.3 How Do Users Compare the Speed and Usefulness of SC3 Sim's Feedback with Traditional Methods of Estimating COA Outcomes?

The SC3 Sim contributed meaningfully to user confidence in plan selection by providing immediate, visualized feedback on COA execution outcomes. Although side-by-side COA comparison was not yet implemented, participants were able to test and revise COAs sequentially, submitting one version, recording the perceived outcomes, and then submitting a modified version for further evaluation. This workflow led users to focus on refining a single seed COA, with survivability (measured by the number of friendly or enemy units surviving the scenario) serving as the dominant metric for plan selection. While this constrained the depth of comparative analysis, the simulator's fast, automatic feedback on outcomes helped clarify the impact of COA changes and increased user trust in the plans they chose to refine further.

The SC3 interface and simulator allowed for reasonably efficient exploration of alternate branches and "what if" scenarios. Participants described the point-and-click interface for tasking and repositioning units as intuitive, with changes reflected quickly in follow-up simulations. This speed and responsiveness supported iterative testing, enabling planners to explore the consequences of tactical adjustments with minimal overhead. In session 3, the integration of COA-GPT added a further layer of flexibility by allowing users to request COA modifications via text prompts. Although some limitations remained in how reliably those modifications were implemented, participants still saw value in being able to regenerate COA content and test alternative approaches with limited friction.

Compared with traditional methods such as wargaming or staff-led COA discussions, participants described the SC3 system as significantly faster and more autonomous. Without the need for facilitator coordination or staff consensus, users could simulate, evaluate, and revise COAs in minutes rather than hours. This represented a substantial reduction in planning friction, especially for individual planners or teams operating under tight time constraints. The interface made key elements of COA execution visible and editable, with overlays for unit positioning, task progression, and terrain features all available in one place. Participants noted that this streamlined access to COA data reduced the need for manual cross-referencing or note-taking.

Although the current system supported early-stage decision making effectively, participants indicated that future versions would benefit from improved tools for comparing multiple COAs, aggregating and organizing tasks at scale, and more robust editing validation to ensure changes are accurately reflected. With these refinements, SC3 Sim and COA-GPT together could support not only rapid scenario testing but also more sophisticated analysis of COA branches, sequels, and doctrinal alignment pushing the ITS toward full-spectrum decision support.

2.5.4 How Well Can SMEs Identify and Trace Back Mission Analysis Outputs to Original Source Documents (e.g., Commander's Intent)? How Effective Is the Interface in Helping Users Access, Trace, and Interpret Mission Analysis Elements Derived from Input Documents?

Most participants were already deeply familiar with the Operation Tropic Tortoise scenario, which reduced the need to refer back to the source documents during early planning. However, when modifying or validating the seed COAs, especially in session 3, SMEs did return to the OPORD and related materials to cross-check decisions and assess consistency with mission guidance. While all participants had access to the original documents and could reason through the likely rationale behind most COA-GPT outputs, there were no embedded features within the interface or COA-GPT outputs that explicitly traced actions or task assignments back to source references, such as a specific paragraph from the commander's intent or a mission task list. This limited the traceability of AI-generated decisions to manual inference.

To better support this need, future iterations of the ITS could incorporate features that explicitly link COA-GPT outputs to the source materials from which they are derived. For example, when COA-GPT assigns a task to a unit like C Company, 1st Squadron, 1st Cavalry Regiment (C 1-1), the system could indicate which parts of the OPORD or mission analysis it drew from to make that assignment.

Additionally, a “why” query interface could allow users to select an element of the COA and ask, for instance, “Why is this unit assigned here?” or “What mission objective is this task supporting?” COA-GPT could then return a rationale grounded in the input materials. Similarly, allowing users to highlight or mark specific elements within the COA and pass them directly into a GPT-powered query, such as “These units appear to be idle, is that intentional?” or “Why aren’t these forces addressing OBJ MONROE?” would significantly improve transparency and trust in the system’s recommendations

2.5.5 Does the SC3 Interface Clearly Present Essential Mission Analysis Elements (e.g., Constraints, Specified/Essential Tasks, Terrain Limitations) in a Way that Supports Early-Stage Planning?

The SC3 interface showed a promising ability to support early-stage planning by providing a clear and accessible environment to visualize and interact with mission analysis components such as unit positioning, terrain features, and initial objectives. SMEs consistently described the system as easy to learn and use, with several praising the map-based UI for grounding planning decisions in known terrain features. The point-and-click interface allowed for relatively quick adjustments to unit tasking, and overlays such as range rings and labeled areas of operation helped to convey key planning elements like support zones or defensible terrain.

However, despite these strengths, there were limitations in how explicitly the interface presented doctrinally critical mission analysis elements. For example, specified and essential tasks were not always clearly identified, and SMEs noted that constraints or timing-based dependencies were not consistently surfaced or linked across views. The absence of a temporal planning structure and explicit phase-based coordination made it difficult to visualize when and how operations should unfold over time. One SME noted that the current interface could not show whether units had fulfilled tasks in proper sequence, such as artillery establishing position before a maneuver force initiated its objective. This feedback suggests that although the interface has a solid visual and interaction foundation, additional features like time-phase tools, doctrinal overlays, and clearer task validation feedback are needed to fully support mission analysis in a realistic, operationally meaningful way.

2.5.6 Did interacting with ITS Clarify Any Mission Planning Considerations?

Interacting with the ITS did not significantly surface new or unexpected mission planning considerations for the participants. This was largely due to the context of

the evaluation: two of the SMEs were already highly familiar with Operation Tropic Tortoise, and the ITS tools were not yet mature enough to generate novel or fully developed COAs that might challenge or expand a user's planning perspective. Although the system offered helpful visualizations and some acceleration of planning tasks, it did not reach a level of output complexity or reasoning depth that would prompt users to rethink or newly consider elements of the mission.

2.5.7 How Frequently Do SMEs Revise COA-GPT's Outputs, and What Specific Deficiencies Motivate Those Changes?

SMEs frequently revised COA-GPT's outputs, primarily to address gaps in task sequencing, operational phasing, and alignment with mission goals. Two of the three participants made substantial edits during session 2. Golf estimated over 50 task changes across 10 units, while Echo reported at least 10 major adjustments. These changes included retasking field artillery to appropriate position areas for artillery, restructuring maneuver elements to ensure proper conditions were set before advancing on objectives, and correcting the timing of unit movements. The primary motivation for these edits was to ensure the generated COAs could realistically achieve the mission objectives. Participants noted that certain actions, such as fixing and seizing operations at key objectives, were represented as independent tasks without adequate attention to timing or dependencies. In some cases, units were directed toward objectives without necessary supporting actions being completed. These findings suggest that while COA-GPT provided a helpful baseline, SMEs still needed to refine and reorganize outputs to meet doctrinal and operational expectations. Future versions of the system could benefit from supporting tools that make such dependencies clearer and streamline refinement across units and phases.

2.6 Experimental Design Lessons

The touchpoint offered a productive opportunity to explore how individual users engage with the SC3 ITS and collect high-value data across usability, task behavior, and subjective feedback. These early insights helped clarify where the system is already supporting key elements of the planning process and where further development is needed. Looking forward, the next phase of experimentation can build on these results by implementing more structured and role-specific testing, particularly to assess how the ITS supports full planning teams. Purposefully designed scenarios with realistic time constraints, operational roles, and shared planning responsibilities will be critical to evaluating how the system affects coordination, task division, and collective decision quality.

As experimentation evolves to include more complex team dynamics and cognitive measures, it will be increasingly important to design studies that intentionally support the use of targeted instrumentation. Tools like BioHarnesses, eye trackers, or smart watches offer great potential for understanding user workload, attention, and stress responses, and they are most informative when aligned with specific phases of an experiment that are designed to elicit the corresponding behaviors. For example, identifying periods of dynamic task load or stable interaction can help ensure the data collected are meaningful and interpretable. Similarly, training protocols should prepare participants not only to use the system, but to do so in ways that reflect the planning expectations of their assigned role. This approach will help ensure that collected data reflects actual decision-making processes rather than system unfamiliarity or procedural artifacts. With careful design, future research can preserve the strengths of prior efforts while advancing toward more rigorous, operationally relevant insights.

2.6.1 MDMP Individual Studies

The touchpoint yielded a valuable set of data across subjective measures, interaction logs, and qualitative feedback, offering insight into both the usability of the ITS and its perceived value across planning roles. These outcomes underscore the success of the instrumentation strategy in capturing user perspectives and system performance. At the same time, the event also revealed areas where future experimentation could be strengthened, particularly in how we prepare individual participants and align data collection tools with specific research questions. As the ITS matures and experimentation shifts toward more operationally realistic settings, the need for role-specific training and targeted instrumentation will become increasingly important.

To draw meaningful conclusions about how individuals use the ITS in a team setting, each participant must engage in the experiment with both a clear understanding of their operational planning role and a sufficient familiarity with the system. This includes structured training on both doctrinal planning responsibilities (e.g., intelligence, logistics, maneuver) and, most importantly for the purpose of this research, ITS tool use. Without this preparation, it becomes difficult to interpret user behavior because it is unclear whether performance reflects actual task dynamics or simple unfamiliarity with the tools. Structured onboarding tailored to individual planning roles, supported by job aids or scenario-specific guides, will help ensure that each participant engages with the system in a manner consistent with operational expectations.

Future experimentation should also deliberately align instrumentation choices with specific research questions, particularly when physiological or cognitive workload

measures are of interest. BioHarnesses, smart watches, or eye trackers can offer valuable insights into user stress, attention, or decision focus, especially when paired with well-designed experimental phases. For example, to evaluate the cognitive impact of dynamic plan changes, we must design scenario segments with meaningful time pressure or unexpected events. Similarly, if eye tracking is used to assess interface usability or visual attention patterns, the task environment must include stable periods where participants are seated and engaging directly with the screen and the setup must support mapping gaze direction to interface features. By ensuring this alignment, experiments may avoid collecting noisy or uninterpretable data as well as avoiding unnecessary constraints on how participants interact or collaborate.

Finally, future experiments should be careful to avoid instrumentation strategies that overburden participants or unintentionally alter natural behavior. This is especially important in team settings, where physical freedom to move, interact, or cluster around a shared display may be central to effective collaboration. If the choice of measurement tools requires participants to remain seated or tethered to a specific device, we risk capturing behavior that reflects experimental constraints rather than operational decision-making. Future experiments should therefore begin with a clear mapping between each research question and the specific data needed to answer it. This approach will support better experimental design, reduce participant burden, and help preserve the ecological validity of the planning environment.

2.6.2 MDMP Team Studies

Although the touchpoint was not structured as a formal team experiment, it provided some limited insight into how the SC3 ITS might support collaborative planning. During session 2, the SMEs participated in a loosely coordinated planning activity that began with a brief discussion of their approach to optimizing a COA and ended with a group review of simulation outcomes. However, the middle phase of the session, when participants were working individually on separate computers, did not involve meaningful team collaboration. Instead, participants experimented independently with the simulator and task editor, attempting to improve outcomes primarily by maximizing unit survivability. In this context, participants openly joked about “gaming the system” rather than refining doctrinally sound or integrated COAs. Though this session did not replicate a real-world staff planning environment, it did surface several lessons for future team-focused experimentation.

The current evaluation could not determine whether the ITS supports alignment of team members’ mental models or enables effective coordination across warfighting

functions. Although participants began with a shared sense of mission intent, there was little indication that they maintained alignment on task sequencing, dependencies, or doctrinal logic. Because there were no assigned planning roles or explicit coordination protocols, it was not possible to assess whether the system helped or hindered shared understanding. That said, SMEs expressed interest in the potential for the ITS to support distributed planning roles, especially if future interface enhancements include tools for assigning task ownership, visualizing interdependencies, and coordinating across functional domains.

To meaningfully evaluate the ITS's impact on team planning, future experimentation must simulate operationally realistic team structures. This includes assigning participants to defined roles such as maneuver, intelligence, fires, and logistics, and ensuring they are trained not just in using the system, but in using it as part of a functioning team. Baseline planning tasks should be conducted using traditional methods to establish current norms of process and outcome quality. Then, comparable scenarios can be executed using the ITS to evaluate differences in performance, coordination, and decision speed. Such structured comparisons are essential to determining the system's value in real-world planning contexts.

Future studies should also include mechanisms to capture communication patterns, planning timelines, and shared knowledge development. Instrumented logging, observational coding, and scenario-based probes can help identify whether the ITS enables more efficient knowledge integration, error detection, and plan adaptation. Scenarios with injected changes in mission conditions or new intelligence will be particularly useful in evaluating whether ITS-supported teams can maintain shared awareness and adapt effectively. Although the present effort could not answer key team-level questions, it highlighted the critical need for properly structured, role-based team experiments to fully evaluate how the ITS supports collaborative planning and decision-making. Similarly, future testing must allow researchers to observe whether—and how—team workflows, role assignments, or responsibilities shift in response to ITS capabilities. For instance, the ability to use natural language to update COAs may enable more centralized editing, potentially reducing the need for distributed task input unless the system is explicitly designed to support it. Likewise, if simulation feedback provides faster answers to key planning questions, teams may adjust how they allocate analytical responsibilities. To study these effects, scenarios must be designed to require interdependent contributions from multiple roles, and the ITS must support features like version tracking, task ownership indicators, and change annotations to assess how the technology shapes collaboration.

2.7 Summarized Self-Report Instrument Scores

This section summarizes the responses to self-report instruments in terms of aggregated scores, broken out by high-level factors where applicable, for instruments that can be scored.

2.7.1 Intake

Table 1 summarizes aggregate scores in response to the questionnaires administered during intake.

Table 1. Summary of scored intake questionnaire responses.

Instrument: factor [scale]	Aggregate ($\mu \pm \sigma$)	Alpha	Echo	Golf
Decision style: rational [1: less rational ↔ 5: more rational]	4.33 ± 0.70	3.6	4.4	5
Decision style: intuitive [1: less intuitive ↔ 5: more intuitive]	2.27 ± 0.31	2.6	2.2	2
Intolerance of Uncertainty: prospective anxiety [1: low anxiety ↔ 5: high anxiety]	2.24 ± 1.11	1	2.57	3.14
Intolerance of Uncertainty: inhibitory anxiety [1: low anxiety ↔ 5: high anxiety]	1.88 ± 0.88	0.92	2.05	2.66
Propensity to trust (LLM agents) [1: low trust ↔ 5: high trust]	3.28 ± 0.51	2.83	3.83	3.17
NASA-TLX (modified) [0: low workload ↔ 100: high workload]	55.80 ± 6.43	59	60	48.4
Team workload [1: low workload ↔ 10: high workload]	5.00 ± 1.20	5	6.2	3.8

2.7.2 Session 1

Table 2 summarizes aggregate scores in response to the questionnaires administered during session 1.

Table 2. Summary of scored session 1 questionnaire responses.

Instrument: factor [scale]	Aggregate ($\mu \pm \sigma$)	Alpha	Echo	Golf
NASA-TLX (modified) [0: low workload ↔ 100: high workload]	54.60 ± 8.15	50.2	64	49.6
Human–computer trust: perceived technical competence [1: low competence ↔ 5: high competence]	3.33 ± 0.70	3.4	2.6	4
Human–computer trust: perceived understandability [1: low understandability ↔ 5: low understandability]	3.53 ± 0.12	3.6	3.4	3.6
Human–computer trust: faith [1: low faith ↔ 5: high faith]	2.47 ± 0.31	2.4	2.2	2.8
UI satisfaction: overall [0: low satisfaction ↔ 100: high satisfaction]	65.40 ± 4.99	59.7	69	67.5
UI satisfaction: screen [0: low satisfaction ↔ 100: high satisfaction]	57.83 ± 11.93	61.5	44.5	67.5
UI satisfaction: terminology and system information [0: low satisfaction ↔ 100: high satisfaction]	76.17 ± 16.82	78.5	58.3	91.7
UI satisfaction: learning [0: low satisfaction ↔ 100: high satisfaction]	70.17 ± 5.91	69.7	64.5	76.3
UI satisfaction: system capabilities [0: low satisfaction ↔ 100: high satisfaction]	82.47 ± 7.57	80.6	76	90.8

2.7.3 Session 2

Table 3 summarizes aggregate scores in response to the questionnaires administered during session 2.

Table 3. Summary of scored session 2 questionnaire responses.

Instrument: factor [scale]	Aggregate ($\mu \pm \sigma$)	Alpha	Echo	Golf
NASA-TLX (modified) [0: low workload ↔ 100: high workload]	44.20 ± 15.51	43.6	60	29
Team workload [1: low workload ↔ 10: high workload]	4.07 ± 1.40	4	5.5	2.7
Team processes: transition processes [1: low functioning ↔ 5: high functioning]	3.40 ± 0.62	3.2	2.9	4.1
Team processes: action processes [1: low functioning ↔ 5: high functioning]	3.07 ± 0.29	3.4	2.9	2.9
Human–computer trust: perceived technical competence [1: low competence ↔ 5: high competence]	3.07 ± 0.81	2.6	2.6	4
Human–computer trust: perceived understandability [1: low understandability ↔ 5: low understandability]	3.60 ± 0.53	4	3	3.8
Human–computer trust: faith [1: low faith ↔ 5: low faith]	2.47 ± 0.50	2	2.4	3
System trustworthiness scale [0: low trustworthiness ↔ 100: high trustworthiness)	65.13 ± 21.03	53.8	52.2	89.4
Technology acceptance model: perceived usefulness [1: low usefulness ↔ 5: high usefulness]	3.93 ± 0.60	4	3.3	4.5
Technology acceptance model: perceived ease of use [1: low ease of use ↔ 5: high ease of use]	3.77 ± 0.84	4.2	2.8	4.3

2.7.4 Session 3

Table 4 summarizes aggregate scores in response to the questionnaires administered during session 3.

Table 4. Summary of scored session 3 questionnaire responses.

Instrument: factor [scale]	Aggregate ($\mu \pm \sigma$)	Alpha	Echo	Golf
NASA-TLX (modified) [0: low workload ↔ 100: high workload]	45.60 ± 12.93	46.6	58	32.2
Human–computer trust: perceived technical competence [1: low competence ↔ 5: high competence]	3.13 ± 0.58	2.8	2.8	3.8
Human–computer trust: perceived understandability [1: low understandability ↔ 5: low understandability]	3.27 ± 0.64	2.8	3	4
Human–computer trust: faith [1: low faith ↔ 5: high faith]	1.87 ± 0.61	2	2.4	1.2

3. Detailed Analysis and Results

This section documents a more in-depth and complete record of the analyses and the results derived from the touchpoint that may not be covered in the high-level findings reported above. This section is organized based on a set of topic-based learning demands we aimed to address with the touchpoint. Callouts to the Appendix are included at the end of the below sections where specific pieces of data inform the corresponding analysis.

3.1 Usability and Features

3.1.1 High-Level ITS Feedback

Across all sessions, SMEs expressed support for the direction of the ITS research and noted that the system is steadily improving. While they agreed the tool is not yet ready to support live COA development, they emphasized that its current capabilities already offer value, especially for individual planners working without a full staff. The system helped users move more quickly through early planning steps and provided a structured starting point for COA development. SMEs saw clear potential in the integration between COA-GPT and the SC3 Sim. Even in its current state, the system enabled users to move from mission analysis to COA sketching and simulation, helping to frame planning decisions. However, several participants encountered mismatches between what COA-GPT said it changed and what appeared in the simulation. These disconnects limited trust in the system and highlighted the need for more consistent and transparent synchronization between planning edits and simulation updates.

The most critical missing feature identified was the lack of support for phasing and task sequencing. SMEs repeatedly noted that without the ability to define actions

over time and across operational phases, it becomes difficult to validate whether a COA is executable. They recommended adding timeline-based tasking, phase-aware visualizations, and simulation playback that allows planners to see how operations unfold step by step. Participants also suggested additions that would support more realistic wargaming. These included incorporating both friendly and enemy COAs, defining evaluation criteria, and enabling comparisons across multiple plan options. SMEs asked for COA-GPT to be expanded so it can propose alternate task organizations or recommend the creation of task forces when appropriate, based on higher-level mission demands.

Looking ahead, SMEs encouraged future touchpoints to include full scenario setups with doctrinal structures, realistic operational constraints, and clear opportunities to test how the ITS supports planning under pressure. One participant summarized the general feeling well: the system is not yet where it needs to be, but it has come a long way and continues to move in the right direction. With continued focus on sequencing, coordination, and output clarity, the ITS has the potential to make planning faster, more consistent, and easier to adapt (Appendix A.1.1.1).

3.1.2 Useful Features

3.1.2.1 What ITS Features Were Useful and How?

SME feedback across both sessions highlighted several ITS features as immediately useful and aligned with operational needs. The most positively received components fell into three major categories: 1) ease of use and intuitive interface, 2) simulation and tasking features, and 3) COA-GPT's document parsing and summarization capabilities.

3.1.2.1.1 Intuitive and accessible UI. Across all sessions, SMEs consistently emphasized that the ITS was easy to learn and use, even in its partially developed state. Participants (Alpha and Golf) specifically noted that the point-and-click interface for directing unit movement was straightforward and efficient, especially when grounded in a terrain-aware map. One SME noted the value of having known terrain features visible, which helped orient planning decisions and simulate real-world operational reasoning. The interface also enabled quick task editing and reassignment using direct interactions rather than complex menus. Alpha described the ease of adding new tasks and moving units, while Echo praised the interface's capacity to visualize COA execution paths and provide immediate status feedback after simulation runs. SMEs also saw potential in the augmented map concept, particularly for enabling shared understanding across distributed command nodes, signaling that the ITS may offer unique advantages in geographically dispersed planning contexts.

3.1.2.1.2 Constructive simulation and plan execution feedback. The simulation component was described as a major strength, particularly for its reusability and responsiveness to changes. Golf highlighted the benefit of being able to repeatedly run the simulation with just minor edits, accelerating iterative COA refinement. Alpha and Echo also highlighted the simulation's ability to quickly reflect the outcomes of input COAs and to automatically display results such as friendly and enemy losses, which enhanced user understanding of potential operational tradeoffs. In particular, the preseeded simulation was valued by participants as a mechanism to avoid manual setup and "drag-and-drop fatigue." Echo appreciated that the baseline scenario already included starting conditions, allowing planners to immediately begin evaluating and modifying courses of action. This strongly supports the hypothesis that automation of scenario initialization is a key value-add of ITS technologies.

3.1.2.1.3 COA-GPT's summarization and planning support. SMEs provided positive commentary on COA-GPT's ability to ingest large amounts of operational data and surface key information quickly. Alpha called this out explicitly, identifying it as valuable for speeding up information processing during the planning phase. Golf also noted that the model continues to improve and already provides plausible and modifiable COA suggestions. This function appears particularly useful for accelerating the initial stages of COA development, providing users with a viable starting point rather than a blank slate.

The integration of COA-GPT with structured tasking workflows was seen as foundational to enabling rapid iteration, and several participants (e.g., Echo) suggested that its outputs made the follow-on simulation more actionable and effective (Appendix A.1.1.2).

3.1.2.2 What Are the Three Most Significant Strengths of COA-GPT Identified by Users?

3.1.2.2.1 Speed and efficiency in parsing and synthesizing operational inputs. Participants praised COA-GPT's ability to quickly absorb and summarize large volumes of mission-related data, including commander's intent, mission tasks, and synchronization requirements. This functionality was highlighted as particularly valuable for accelerating the early stages of COA development. One SME noted that COA-GPT's output offered "a good start" and saved considerable time compared with starting from a blank slate. The system was described as helpful in transforming mission analysis into an initial structured plan with minimal user prompting, enabling individual planners to move more quickly into task organization and refinement.

3.1.2.2.2 Natural language-based plan modification. The ability to use natural language instructions to iteratively refine COAs was another standout strength. In session 3, participants interacted with COA-GPT to make phase-specific edits to the seed COAs, such as repositioning units or adjusting their tasking sequences. Although the system did not always execute instructions flawlessly, SMEs expressed that the capacity to prompt changes like “move this unit after the breach” or “add artillery prep before maneuver” represented a promising and intuitive interaction mode. This kind of interface allowed them to engage in the planning process in a manner that mirrored natural staff communication, reducing the cognitive burden of manipulating plans through menus or abstract inputs.

3.1.2.2.3 Support for generating structured, phase-based plans. SMEs recognized COA-GPT’s emerging capability to structure plans by temporal phases and associate unit actions with specific points in the operation timeline. While this functionality was still under development and not always fully synchronized with the simulator, users appreciated that COA-GPT could at least begin to build phased plans, distinguishing between early shaping operations and main efforts. This structuring was seen as critical for operational realism and mission execution, as it helps planners track dependencies, establish decision points, and ensure doctrinal alignment across phases.

3.1.3 Problematic Features

3.1.3.1 What ITS Features Were Problematic and How?

SMEs were generally positive about the ITS and its potential to support planning workflows, but they also identified several recurring usability limitations and capability gaps. Feedback focused on three major problem areas: 1) limitations in COA-GPT accuracy and planning logic, 2) challenges with simulation structure and COA comparison, and 3) friction in the UI for planning coordination and clarity.

3.1.3.1.1 COA-GPT limitations in terrain awareness and planning validity. Across both sessions, SMEs noted that COA-GPT did not adequately account for terrain constraints or doctrinal priorities in its generated COAs. Alpha highlighted that its current inability to incorporate terrain meaningfully limits its usefulness in realistic COA generation. Golf stated that the COAs lacked viability at the COL level and missed key operational objectives: for instance, omitting OBJ MONROE, which was identified as the division’s main effort. This undermined SME trust in the tool and its potential for producing plausible plans under realistic conditions.

3.1.3.1.2 Simulation structure and COA comparison limitations. The constructive simulation was praised for its potential, but current implementation lacked several

features critical for operational planning. Alpha noted the inability to coordinate actions across phases or visualize operations over time, which made it difficult to assess temporal synchronization. Both Echo and Golf flagged the lack of persistent COA comparison tools as a major issue. Users could not easily save, reference, or compare COAs within the interface, requiring external note-taking and making iterative refinement cumbersome. Echo also noted that COAs did not always satisfy critical directives from the OPORD, compounding trust concerns.

3.1.3.1.3 UI complexity and control friction. Numerous UI-level challenges were raised across both sessions. SMEs noted that task insertion and unit-task alignment were often confusing or restrictive (Alpha), with issues like the system resetting task locations to the unit position rather than a prior task. Echo and Golf emphasized that the system became harder to navigate as the number of units increased, and that the lack of aggregation mechanisms (e.g., moving a battalion as a whole) created unnecessary friction for formation-level planning. Display clutter was another concern: SMEs asked for ways to toggle off visual elements like movement paths and range rings that interfered with focus. Additionally, default UI behaviors (such as all layers reappearing after navigating away from a view) created repetitive cleanup work and reduced efficiency (Golf).

Across the board, SMEs conveyed that many of these issues did not reflect a fundamental design flaw but rather areas in need of refinement, prioritization, and increased configurability. The presence of these friction points did not negate the system's value, but did constrain its current effectiveness, especially for more complex planning tasks or dense operational scenarios (Appendix A.1.1.3).

3.1.3.2 What Are the Three Most Significant Areas of COA-GPT for Improvement Identified by Users?

3.1.3.2.1 Poor integration of terrain and doctrinal constraints. Across both qualitative comments and numerical ratings, SMEs consistently cited COA-GPT's failure to meaningfully incorporate terrain features and doctrinal planning principles. Users reported that COAs often ignored key spatial boundaries like phase lines, omitted preconditions for maneuver actions, and failed to align sequencing with standard combined arms practices. For instance, one SME noted that COA-GPT recommended movement toward OBJ MONROE without first achieving preconditions at OBJ MADISON, which is an error that contradicts operational logic. Others highlighted that artillery units were not correctly positioned before being expected to support maneuver units, violating task dependency rules. These issues eroded trust in COA-GPT's ability to generate executable, mission-valid plans and suggest a need for tighter integration of terrain models, doctrinal rules, and constraint checks into the generation process.

3.1.3.2.2 Inconsistent responsiveness to natural language edits. Although natural language interaction is one of COA-GPT's standout positive features, its current implementation was viewed as somewhat unreliable and unpredictable. Participants in session 3 attempted to use natural language to adjust unit assignments, change task timing, and add phases, but sometimes found that the model either failed to implement the change or reported success without making visible adjustments. For example, Echo reported that COA-GPT deleted all phases except Phase 1 after a prompt and failed to fully reimplement the missing phases when prompted again. Another user (Alpha) noted that while edits to 1-1CAV and 5-20IN were initially successful, subsequent changes to 1-6IN not only failed but disrupted unrelated units' plans. These issues suggest the need for better prompt parsing, validation, and confirmation feedback loops such as displaying a summary of changes made or requesting clarification when ambiguity is detected.

3.1.3.2.3 Low reliability in task sequencing and synchronization. Participants found that COA-GPT repeated tasks across some phases, skipped critical handoffs, or misaligned task execution with mission timelines. Alpha explicitly noted that some tasks were repeated across phases when they should have transitioned, and Echo described cases where task assignments were incompletely or incorrectly reflected in the synchronization matrix. This reduced user confidence in COA-GPT's ability to maintain logical progression and interunit coordination across phases. These deficiencies were reflected in low average scores on items assessing time-based phase management (2.0/6) and alignment with mission requirements (2.66/6), pointing to a systemic issue in temporal reasoning and task handover representation.

3.1.4 Lacking Features

SMEs were asked "Which element of the ITS is most lacking, specifically in terms of hindering your ability to conduct planning in a sufficiently realistic manner (e.g., gaps in representations, oversimplifications, inaccuracies)? What would you say are the most critical of these lacking areas to address? In what way would you like them addressed?"

The most significant limitation of the ITS that was reported was the lack of support for temporal sequencing and phase-based coordination. Without the ability to break operations into time-based phases or define task dependencies, the system prevents realistic modeling of how units interact over time. SMEs explained that this limitation undermines the doctrinal validity of the planning process. For instance, in one scenario, a brigade was shown seizing a key objective before supporting artillery units had moved into position, violating standard sequencing. This made it difficult for users to determine whether COAs were actually executable. SMEs

recommended integrating a timeline-based planning structure that allows tasks to be assigned by phase, ordered by dependency, and visualized with clear operational timelines. They also requested simulation playback tools aligned with these phases, so that planners can evaluate branch points and decision triggers in context.

A second major issue was COA-GPT's inability to account for terrain, operational priorities, or unit constraints in a consistent and realistic way. Several SMEs noted that the generated COAs ignored critical objectives or proposed implausible unit taskings. For example, one COA failed to mention OBJ MONROE, the division's main effort, casting doubt on the utility of the output. While SMEs appreciated COA-GPT's ability to summarize orders and generate structure, they emphasized the need for better grounding in mission inputs and more transparent justification for its recommendations. They advised incorporating checks for doctrinal completeness, ensuring that all specified and implied tasks are addressed, and improving integration with terrain overlays and order references.

The UI also presented challenges, especially at scale. SMEs described cluttered displays, difficulty tracking unit status, and inefficient controls when working with large formations. Some expressed frustration at having to move units individually rather than as a combined element and not being able to filter out irrelevant visual elements like routes from nonessential units. Suggested improvements included tools for grouping units by command structure, toggling overlays, searching for specific units, and quickly switching between COAs or planning perspectives.

In short, while the ITS shows strong potential, several key limitations are currently preventing it from supporting realistic, scalable planning. Addressing these issues through improved temporal structuring, grounded and traceable COA generation, and a more usable and scalable interface would significantly improve its ability to serve as a reliable planning tool (Appendix A.1.1.4).

3.1.5 Planning Effects

3.1.5.1 What Impact Do You Think ITS Had on Your Planning Process?

Although the ITS was not evaluated under conditions equivalent to a full staff COA development process, participants indicated that it still offered meaningful support for individual planners during early phases of mission planning. The combined use of COA-GPT and the SC3 Sim helped users move through key planning steps more quickly than they otherwise might on their own, especially by providing a structured starting point for COA development. One SME noted that they were able to reach multiple MDMP phases more rapidly with ITS support, even though they were working without the benefit of a full planning team.

The ability to generate and test an initial COA sketch through the linked modules of COA-GPT and the simulator was seen as a valuable foundation. SMEs appreciated that COA-GPT could synthesize mission inputs and propose an initial COA, and that the simulator could provide a basic operational run-through of the plan. However, several limitations emerged in how useful and trustworthy those outputs were for deeper planning. For example, one SME described situations where COA-GPT claimed to have completed a task that, on review, was not accurately reflected in the plan. Others encountered challenges with editing COAs and ensuring that updates were faithfully implemented within the system. These issues reduced confidence in the outputs, especially in terms of how well they would hold up under scrutiny in a collaborative or operational setting.

Despite these concerns, the ITS was widely described as a promising and rapidly maturing tool. Participants agreed that, for individual planners working in time-constrained or resource-limited contexts, the system already offers a valuable head start and a clear structure for further refinement. One SME summarized this sentiment by noting that while the ITS is not yet capable of replacing full traditional staff processes, it is already a strong starting point and has made significant progress. With continued development focused on improving output accuracy, task traceability, and editing responsiveness, the ITS has the potential to meaningfully accelerate and support COA planning both for individuals and, eventually, for full teams.

3.1.5.2 How Well Did ITS Integrate Terrain, Unit Capabilities, and Mission Objectives? How Well Did the Constructive Simulation Component Do This? How Well Did the COA-GPT Component Do This? Do SMEs Perceive the Terrain Analysis and Unit Deployment Within COA-GPT Outputs to Align with Mission Requirements and Constraints?

The ITS demonstrated a foundational ability to integrate terrain, unit capabilities, and mission objectives, with particularly strong contributions from the simulation component. The SC3 Sim provided layered visualizations of terrain, mission graphics, and unit positions, which SMEs found useful for orienting planning and understanding operational constraints. Terrain features were clearly labeled and supported situational awareness, helping users interpret how unit movements might be influenced by environmental factors. SMEs noted that this supported intuitive interactions with the mission space, though they also expressed a desire for more flexible graphic control measures, clearer phase line integration, and additional tools for designating relevant areas such as no-fire zones or control boundaries.

The constructive simulation did well in modeling unit properties and showing mission objectives in relation to terrain, offering automatic outcomes such as loss

estimates and unit status changes after COA execution. Participants found these feedback features helpful, particularly when evaluating seeded scenarios that avoided time-consuming manual setup. However, they also indicated that the simulation lacked support for temporal coordination and did not always reflect doctrinal sequencing, such as clearing phase areas prior to advancing. While the simulation effectively linked unit capabilities with terrain visualization, its realism was limited by the absence of operational phasing, support-to-main effort relationships, and mission dependencies.

COA-GPT's integration of terrain and mission elements was more limited. Although it performed well in parsing mission inputs and generating initial COA sketches aligned with problem and mission statements, SMEs noted that its outputs sometimes overlooked essential tasks or failed to reflect key terrain considerations. For instance, multiple participants reported that important objectives like OBJ MONROE were omitted, and that COA-GPT occasionally returned plans that were infeasible based on terrain or unit limitations. Overall, while COA-GPT provided a strong starting point and supported faster mission analysis, its current integration with the physical and operational environment remains an area for continued refinement. SMEs suggested that improved traceability, doctrinal validation, and terrain-awareness features would significantly enhance its utility in realistic planning.

3.1.5.3 How Well Can SMEs Identify and Trace Back Mission Analysis Outputs to Original Source Documents (e.g., Commander's Intent)? How Effective Is the Interface in Helping Users Access, Trace, and Interpret Mission Analysis Elements Derived from Input Documents?

Most participants were already deeply familiar with the Operation Tropic Tortoise scenario, which reduced the need to refer back to the source documents during early planning. However, when modifying or validating the seed COAs, especially in session 3, SMEs did return to the OPORD and related materials to cross-check decisions and assess consistency with mission guidance. While all participants had access to the original documents and could reason through the likely rationale behind most COA-GPT outputs, there were no embedded features within the interface or COA-GPT outputs that explicitly traced actions or task assignments back to source references, such as a specific paragraph from the commander's intent or a mission task list. This limited the traceability of AI-generated decisions to manual inference.

To better support this need, future iterations of the ITS could incorporate features that explicitly link COA-GPT outputs to the source materials from which they are derived. For example, when COA-GPT assigns a task to a unit like C 1-1, the

system could indicate which parts of the OPORD or mission analysis it drew from to make that assignment. Additionally, a “why” query interface could allow users to select an element of the COA and ask, for instance, “Why is this unit assigned here?” or “What mission objective is this task supporting?” COA-GPT could then return a rationale grounded in the input materials. Similarly, allowing users to highlight or mark specific elements within the COA and pass them directly into a GPT-powered query, such as “These units appear to be idle, is that intentional?” or “Why aren’t these forces addressing OBJ MONROE?” would significantly improve transparency and trust in the system’s recommendations

3.1.5.4 Does the SC3 Interface Clearly Present Essential Mission Analysis Elements (e.g., Constraints, Specified/Essential Tasks, Terrain Limitations) in a Way that Supports Early-Stage Planning?

The SC3 interface showed a promising ability to support early-stage planning by providing a clear and accessible environment to visualize and interact with mission analysis components such as unit positioning, terrain features, and initial objectives. SMEs consistently described the system as easy to learn and use, with several praising the map-based UI for grounding planning decisions in known terrain features. The point-and-click interface allowed for relatively quick adjustments to unit tasking, and overlays such as range rings and labeled areas of operation helped to convey key planning elements like support zones or defensible terrain.

However, despite these strengths, there were limitations in how explicitly the interface presented doctrinally critical mission analysis elements. For example, specified and essential tasks were not always clearly identified, and SMEs noted that constraints or timing-based dependencies were not consistently surfaced or linked across views. The absence of a temporal planning structure and explicit phase-based coordination made it difficult to visualize when and how operations should unfold over time. One SME noted that the current interface could not show whether units had fulfilled tasks in proper sequence, such as artillery establishing position before a maneuver force initiated its objective. This feedback suggests that although the interface has a solid visual and interaction foundation, additional features like time-phase tools, doctrinal overlays, and clearer task validation feedback are needed to fully support mission analysis in a realistic, operationally meaningful way.

3.1.5.5 Did Interacting with ITS Clarify any Mission Planning Considerations You Had not Thought of?

Interacting with the ITS did not significantly surface new or unexpected mission planning considerations for the participants. This was largely due to the context of the evaluation: two of the SMEs were already highly familiar with Operation Tropic Tortoise, and the ITS tools were not yet mature enough to generate novel or fully developed COAs that might challenge or expand a user's planning perspective. Although the system offered helpful visualizations and some acceleration of planning tasks, it did not reach a level of output complexity or reasoning depth that would prompt users to rethink or newly consider elements of the mission.

3.1.6 COA Branches, What Ifs, and Comparisons

3.1.6.1 How Does the Ability to Visualize Outcomes of Different COAs in the SC3 Sim Affect User Confidence in Plan Selection? Does the Interface Allow Users to Effectively Modify COAs and Explore Alternate Branches or Sequels Based on New Insights? Do Users Feel the Interface Allows Them to Explore "What If" Scenarios with Sufficient Speed and Fidelity to Support Decision Cycles? How Well Does the SC3 Sim Support Iterative Refinement of COAs by Enabling Comparison of Projected Outcomes? How Do Users Compare the Speed and Usefulness of SC3 Sim's Feedback to Traditional Methods of Estimating COA Outcomes (e.g., Wargaming or Staff Discussion)? To What Extent Does the Interface Reduce Planning Friction by Streamlining Access to Essential Data and COA Outputs?

The SC3 Sim contributed meaningfully to user confidence in plan selection by providing immediate, visualized feedback on COA execution outcomes. Although side-by-side COA comparison was not yet implemented, participants were able to test and revise COAs sequentially, submitting one version, recording the perceived outcomes, and then submitting a modified version for further evaluation. This workflow led users to focus on refining a single seed COA, with survivability (measured by the number of friendly or enemy units surviving the scenario) serving as the dominant metric for plan selection. While this constrained the depth of comparative analysis, the simulator's fast, automatic feedback on outcomes helped clarify the impact of COA changes and increased user trust in the plans they chose to refine further.

The SC3 interface and simulator allowed for reasonably efficient exploration of alternate branches and "what if" scenarios. Participants described the point-and-click interface for tasking and repositioning units as intuitive, with changes reflected quickly in follow-up simulations. This speed and responsiveness

supported iterative testing, enabling planners to explore the consequences of tactical adjustments with minimal overhead. In session 3, the integration of COA-GPT added a further layer of flexibility by allowing users to request COA modifications via text prompts. Although some limitations remained in how reliably those modifications were implemented, participants still saw value in being able to regenerate COA content and test alternative approaches with limited friction.

Compared with traditional methods such as wargaming or staff-led COA discussions, participants described the SC3 system as significantly faster and more autonomous. Without the need for facilitator coordination or staff consensus, users could simulate, evaluate, and revise COAs in minutes rather than hours. This represented a substantial reduction in planning friction, especially for individual planners or teams operating under tight time constraints. The interface made key elements of COA execution visible and editable, with overlays for unit positioning, task progression, and terrain features all available in one place. Participants noted that this streamlined access to COA data reduced the need for manual cross-referencing or note-taking.

Although the current system supported early-stage decision making effectively, participants indicated that future versions would benefit from improved tools for comparing multiple COAs, aggregating and organizing tasks at scale, and more robust editing validation to ensure changes are accurately reflected. With these refinements, SC3 Sim and COA-GPT together could support not only rapid scenario testing but also more sophisticated analysis of COA branches, sequels, and doctrinal alignment, pushing the ITS toward full-spectrum decision support.

3.1.6.2 How Satisfied Are Users with COA-GPTs Performance in Facilitating MDMP? How Does COA-GPT Impact the Speed, Clarity, or Confidence of Decisions Compared with Traditional MDMP Workflows? Do Users Find COA-GPT Easier or Harder to Use than Current Planning Tools?

COA-GPT was broadly viewed by participants as a valuable asset for accelerating the early phases of the MDMP, especially in its ability to generate an initial COA from complex input data. One SME described it as offering the benefit of “starting with something on the page,” which significantly reduced the time and effort typically required to parse operational orders and convert them into actionable plans. Rather than beginning with a blank slate, users were able to quickly move into reviewing and refining a base plan, which provided momentum to the planning process. Another major benefit of COA-GPT was its capacity to digest dense planning inputs and present a coherent output that aligned, at least in structure, with expected operational phases and doctrinal components. Participants noted that the

model could identify key mission elements and propose plausible unit assignments and synchronization. This made it easier for users to conduct initial planning and orientation activities, particularly without a full staff. The ability to use natural language for prompting and revision was also seen as promising, supporting a more intuitive and accessible form of interaction.

Despite these strengths, current system limitations prevent COA-GPT from fully realizing its potential as an MDMP accelerator. Users reported multiple instances where the tool failed to make intended edits, misunderstood the scope of requested changes, or misrepresented those changes in the simulation. These issues affected user confidence, especially when outputs required manual inspection or re-implementation to ensure accuracy.

Although COA-GPT is not yet ready to replace traditional planning tools, it is already supporting faster early-stage COA development and improved user engagement. Participants agreed that with improvements to edit handling, constraint awareness, and traceability, COA-GPT could become substantially easier to use and more effective than current systems. With continued development, it has the potential to reduce planning time and help teams produce higher quality COAs (Appendix A.1.1.5).

3.1.7 Trust and Reliance

3.1.7.1 What Level of Confidence Do Human Analysts and Commanders Have in the Outputs Provided by COA-GPT? Do Users Trust the Simulation's Projected Results as Valid Reflections of Operational Dynamics, or Do They Require Manual Verification? What Interface or Visualization Features Contribute Most to User Trust in COA-GPT Versus Analog?

Based on the survey results and supporting feedback in the trust evaluation documents, user confidence in the SC3 ITS remains cautious but constructive. Across both human-computer trust scales and system trustworthiness ratings, participants conveyed moderate levels of trust in the system's underlying logic and displayed a practical understanding of how to use it for planning decisions. However, the responses also reflected clear limits in confidence, particularly when it came to accepting COA-GPT outputs without additional verification or scrutiny.

3.1.7.1.1 confidence in COA-GPT outputs. Although participants agreed that COA-GPT often used the information they entered correctly and could generate viable planning content, they stopped short of affirming that its outputs matched the quality of a highly competent human planner. Trust scores related to COA-GPT's problem knowledge and decision logic averaged approximately 3 out of 5,

suggesting that while the system is seen as competent, its authority is not yet equivalent to that of a trusted human staff officer. Ratings on items such as “the system provides advice as good as a highly competent person” and “I believe the system over myself when I am uncertain” were notably lower (average ratings between 2.3 and 2.7), underscoring the need for human oversight. Participants reported that while they could often follow what the system was doing and understood how to use it for decisions, they still felt uncertain about whether it had properly reasoned through its recommendations. The ability to understand and predict system behavior was rated moderately (scores generally in the low- to mid-60s on a 100-point scale), with variability suggesting uneven clarity across different system interactions.

3.1.7.1.2 Confidence in simulation outputs and need for verification. Trust in the SC3 Sim was higher than in COA-GPT, particularly regarding its consistency and reliability. For example, the item “the system responds similarly in similar circumstances” received one of the highest trust scores across all metrics (mean 86.3). This consistency was reassuring to participants and supported confidence in iterative plan refinement. However, the overall trust in simulation outputs still did not fully negate the need for manual verification. Feedback indicated that while users valued the ability to visualize COA outcomes (e.g., through loss rates and terrain-based movement), they still relied on their own judgment to interpret simulation results. This was especially true when COA-GPT generated plans that omitted critical objectives or violated sequencing expectations. The simulation’s visual playback helped clarify what was happening, but participants remained alert to discrepancies and did not treat the system’s assessment as final.

3.1.7.1.3 Interface and visualization features supporting trust. Positively rated items included the ability to visualize unit movement, toggle task assignments within the map view, and receive immediate feedback on COA execution. Participants noted that the predictability of unit behavior and simulation replay functions improved their sense of system reliability. Despite these strengths, feedback also emphasized the importance of future design changes to further support trust. Participants recommended integrating features that allow traceability from COA elements back to source documents, explaining how specific decisions were derived. For instance, they wanted the ability to click on a tasking and query why it was assigned, or see which elements of the commander’s intent or mission analysis informed a particular COA choice.

3.1.7.2 How High Is the Trust Factor of the COA-GPT–Provided Information (Does COA-GPT Include Footnotes for Source Identification)?

Participant feedback and survey responses suggest that users view COA-GPT as a promising planning support tool, though one that still requires refinement before it can fully earn user trust in high-stakes operational contexts. On average, SMEs rated their trust in COA-GPT’s outputs around the midpoint across scales, indicating cautious optimism. While the tool was not yet considered fully reliable for doctrinally critical decisions, participants acknowledged its usefulness in rapidly generating baseline COAs that could be refined and adapted through user input.

SMEs noted that COA-GPT often provided a helpful starting point that accelerated early planning phases. However, confidence in its outputs varied depending on the complexity of the scenario and the degree of accuracy required. When asked whether the model accurately incorporated their inputs or responded correctly in edge-case conditions, participants offered more reserved ratings. This suggests that while COA-GPT shows strong potential, users still felt the need to verify its outputs manually, particularly in cases involving complex sequencing, terrain limitations, or nuanced mission requirements.

Despite these caveats, participants were encouraged by the system’s trajectory and acknowledged that COA-GPT already contributes meaningfully to the planning process. Its natural language interface and ability to synthesize orders into coherent COA sketches were viewed as core strengths. When used in conjunction with the simulation tools for iterative refinement, the outputs became more actionable and easier to evaluate. Participants emphasized that with improved responsiveness to user edits, stronger incorporation of doctrinal guidance, and enhanced transparency around how recommendations are generated, COA-GPT could become a highly trusted asset.

In summary, user confidence in COA-GPT is cautiously optimistic. Although it is not yet positioned to replace the judgment of trained staff, COA-GPT has already demonstrated value as a tool to accelerate planning workflows and support initial COA development. With further improvement to its accuracy, traceability, and user feedback mechanisms, COA-GPT has the potential to become a trusted and integral part of operational planning (Appendix A.1.1.6).

3.2 Team Performance

3.2.1 Communication and Awareness

Questions posed included the following:

- Do team members exhibit alignment in their mental models of the plan, terrain, and threat environment after leveraging the technologies?
- Do teams share critical assumptions and information when using the system, or does reliance on the technology suppress some forms of communication?
- Do the technologies improve the team's ability to detect inconsistencies or errors in COA logic or assumptions?
- Are teams able to identify, share, and integrate new knowledge during planning as COA iterations and tests evolve?
- How well does the team maintain shared situation awareness in the presence of rapidly changing simulated outputs?

Although the touchpoint was not structured as a formal team experiment, it provided some limited insight into how the ITS might support collaborative planning. During session 2, the SMEs participated in a loosely coordinated planning activity that began with a brief discussion of their approach to optimizing a COA and ended with a group review of simulation outcomes. However, the middle phase of the session, when participants were working individually on separate computers, did not involve meaningful team collaboration. Instead, participants experimented independently with the simulator and task editor, attempting to improve outcomes primarily by maximizing unit survivability. In this context, participants openly joked about “gaming the system” rather than refining doctrinally sound or integrated COAs. Though this session did not replicate a real-world staff planning environment, it did surface several lessons for future team-focused experimentation.

The current evaluation could not determine whether the ITS supports alignment of team members' mental models or enables effective coordination across warfighting functions. Although participants began with a shared sense of mission intent, there was little indication that they maintained alignment on task sequencing, dependencies, or doctrinal logic. Because there were no assigned planning roles or explicit coordination protocols, it was not possible to assess whether the system helped or hindered shared understanding. That said, SMEs expressed interest in the potential for the ITS to support distributed planning roles, especially if future interface enhancements include tools for assigning task ownership, visualizing interdependencies, and coordinating across functional domains.

To meaningfully evaluate the ITS's impact on team planning, future experimentation must simulate operationally realistic team structures. This includes assigning participants to defined roles such as maneuver, intelligence, fires, and logistics, and ensuring they are trained not just in using the system, but in using it as part of a functioning team. Baseline planning tasks should be conducted using

traditional methods to establish current norms of process and outcome quality. Then, comparable scenarios can be executed using the ITS to evaluate differences in performance, coordination, and decision speed. Such structured comparisons are essential to determining the system's value in real-world planning contexts.

Future studies should also include mechanisms to capture communication patterns, planning timelines, and shared knowledge development. Instrumented logging, observational coding, and scenario-based probes can help identify whether the ITS enables more efficient knowledge integration, error detection, and plan adaptation. Scenarios with injected changes in mission conditions or new intelligence will be particularly useful in evaluating whether ITS-supported teams can maintain shared awareness and adapt effectively. Although the present effort could not answer key team-level questions, it highlighted the critical need for properly structured, role-based team experiments to fully evaluate how the ITS supports collaborative planning and decision-making.

3.2.2 Technology Teammate Perception

Questions posed included the following:

- Do team members conceptualize the AI powered technologies as a subordinate, peer, or superior entity in the planning hierarchy?
- Are there differences in how team members delegate cognitive responsibility to COA-GPT (e.g., information filtering, reasoning, prioritization)?
- What are the effects of COA-GPT and the SC3 simulator and interface tools on team performance, cognition, and coordination?
- How does team trust in COA-GPT vary across members, and how does that impact consensus or conflict?

Although the touchpoint sessions were not structured to directly answer questions about team cognition or hierarchical perceptions of AI systems, the survey data and participant behaviors still offer some early insight into how users may relate to COA-GPT in a team setting. Responses suggest that users conceptualize the system in varying ways. One participant described it as a helpful tool for quickly processing planning inputs, while another emphasized its limitations and the need to cross-check its outputs. These differences imply that users may not yet share a stable mental model of COA-GPT's role in planning, and that further work is needed to clarify how the system is framed whether as a subordinate assistant, a peer contributor, or an automated decision-maker. Participants also appeared to delegate cognitive responsibility to COA-GPT in different ways. Some leaned on it to generate complete COAs and only intervened to adjust key units or fix apparent

omissions. Others took a more hands-on approach, modifying tasking and sequencing directly based on their own doctrinal expectations. This variation suggests that future experimentation should examine how users offload different types of reasoning to the AI, such as identifying implied tasks, managing timelines, or organizing units by echelon. Understanding this delegation behavior will be essential for refining both the system's functionality and its transparency mechanisms.

Even though team coordination was not a formal focus of this event, the individual survey responses revealed differing levels of trust in COA-GPT's output quality and reasoning. These differences are important because uneven trust in AI recommendations could complicate consensus-building in future team settings. If one team member accepts a generated COA at face value while another doubts its viability, disagreement may arise over whether and how to refine the plan. Future research should examine how such trust asymmetries influence collaborative planning, especially under time constraints or when high-stakes decisions are involved.

To move forward, future studies must include real teams engaged in cooperative planning tasks under both traditional and ITS-supported conditions. Only through structured team trials with observable role-based interactions, decision cycles, and planning iterations can we evaluate how AI tools like COA-GPT affect team cognition, coordination, and confidence. These studies should incorporate measures of shared understanding, cross-role communication, and consensus formation to assess whether AI assistance enables or inhibits collaborative planning.

3.2.3 Process and Coordination

Questions posed included the following:

- How does the use of the technologies influence the sequencing and timing of planning activities within the team?
- Do teams revise their internal workflows, role assignments, or task division strategies when working with the technologies?
- Does the use of the technologies facilitate or hinder cross-understanding of team roles and responsibilities during mission planning?

The touchpoint was not structured in a way that enabled meaningful evaluation of how the ITS influences the sequencing and timing of planning activities within a team. Participants did not operate with formalized roles or coordinated timelines, and the primary focus was on exploring the capabilities of the system from an individual planning perspective. As a result, we were unable to assess how ITS

tools such as COA-GPT or the SC3 Sim might shift the natural flow of collaborative planning tasks or decision milestones.

To meaningfully answer these questions in the future, experimentation should involve teams with clearly assigned roles and responsibilities that reflect operational command post structures. Teams should complete planning tasks both with and without the ITS, allowing researchers to observe whether the technology compresses, reorders, or extends traditional planning phases. This includes examining whether certain phases such as mission analysis, COA development, or wargaming become more tightly coupled or are reordered when technology accelerates early tasking or feedback. Capturing these dynamics will require synchronized behavioral logging, time-based task segmentation, and potentially eye-tracking or interface interaction analysis to trace the path of attention and coordination.

Similarly, future testing must allow researchers to observe whether and how team workflows, role assignments, or responsibilities shift in response to ITS capabilities. For instance, the ability to use natural language to update COAs may enable more centralized editing, potentially reducing the need for distributed task input unless the system is explicitly designed to support it. Likewise, if simulation feedback provides faster answers to key planning questions, teams may adjust how they allocate analytical responsibilities. To study these effects, scenarios must be designed to require interdependent contributions from multiple roles, and the ITS must support features like version tracking, task ownership indicators, and change annotations to surface how the technology shapes collaboration.

3.2.4 Cognitive Load

Questions posed included the following: How does the use of the technology affect collective cognitive workload?

This factor was not directly tested at the touchpoint, as there was no traditional or baseline planning condition for comparison. Although individual and group workload data were collected using NASA-TLX and Group Workload scales, the limited number of participants and lack of consistent role assignments or team coordination meant that the resulting scores were difficult to interpret. Across sessions, reported workload ratings were generally moderate and did not vary meaningfully between interactions with COA-GPT, the SC3 Sim, or the combined ITS. However, this lack of differentiation is likely attributable to the unstructured nature of the activity rather than to the systems' impact on workload. Participants primarily operated independently or in loosely coordinated ways, and task demands were not distributed across functional roles. As a result, any cognitive load

associated with planning was borne individually, and there was little opportunity to observe how the technologies might influence shared workload dynamics, division of labor, or coordination overhead. In a fully staffed team setting, workload impacts might differ substantially, depending on how the system is used to centralize, automate, or distribute different elements of the planning process.

Future research should focus on designing comparative studies that examine cognitive workload under both traditional and ITS-supported conditions using fully formed teams. These teams should perform structured planning tasks with clearly assigned roles (e.g., maneuver, fires, logistics, intelligence), allowing researchers to assess how the technologies impact both individual and collective workload. Data collection should include not only workload self-reports but also behavioral markers of overload, task switching, and communication efficiency. Doing so will provide a clearer picture of whether and how ITS capabilities reduce planning burden, redistribute effort across the team, or introduce new sources of cognitive strain.

3.2.5 Decision-Making

Questions posed included the following: Are teams with higher initial trust in AI more likely to defer decisions to COA-GPT outputs?

This question could not be meaningfully answered within the context of the touchpoint due to several limitations in study design. Most notably, there was only a single small group of participants rather than multiple teams, and those participants did not function as a true planning team with sustained collaboration or decision-making under shared conditions. Although participants completed individual difference measures related to AI trust (such as the LLM propensity to trust, uncertainty tolerance, and intuitive decision-making scales), the narrow participant pool and limited variation across responses made it impossible to meaningfully evaluate the relationship between trust dispositions and observable behaviors. Additionally, the nature of the session activities did not require or encourage participants to defer decisions to COA-GPT outputs. Instead, participants largely treated COA-GPT as a tool to generate initial drafts or starting points for plans, with significant manual revision and skepticism applied to those outputs. There was no structured team consensus process that would allow researchers to observe whether participants with higher trust traits deferred more readily to AI outputs, nor were there conflicting recommendations or decisions that might provoke deference versus contestation behaviors.

To explore this question in future research, a more robust experimental design is required. Multiple teams should be formed and evaluated, with variation in team

composition and individual trust traits. These teams should face planning tasks that include ambiguous or contested decision points, where reliance on AI-generated outputs versus human judgment becomes a salient choice. Observing when and why different teams or individuals choose to defer to COA-GPT and how that shapes planning quality, team dynamics, and confidence will provide critical insight into how AI trust traits interact with real-world decision-making under collaborative conditions.

Session 2 was the teaming session, where participants were asked to work together on COA generation and assessment. After this session they completed several questionnaires.

The Team Processes Scale (Mathieu et al. 2020) is a 50-item questionnaire that assesses team processes by evaluating 10 first-order constructs that group onto 3 second-order constructs (i.e., transition processes, action processes, interpersonal processes). We have reduced the items to 18 in this study, assessing 2 second-order constructs: transition and action processes. Each item is scored using five-point Likert scales where 1 = not at all and 5 = to a very great extent. Higher scores indicate better functioning team processes and lower scores indicate areas that could be improved (Table 5).

“Transition processes” are behaviors where the team reviews their previous efforts, interprets their environment, and prepares for future work. The consensus among the SMEs appears to be they were at or above average successful in their transition processes.

“Action processes” are behaviors that team members engage in while they are working towards goal accomplishment. The SMEs evaluated their coordination efforts as successful but did not report their efforts on interpreting and communicating necessary information as successful. Their assessment as to how well they did as a team tracking team resources was mixed but still ranged from average to above average (Appendix A.2).

Table 5. Session 2 team process scale summary.

Second-order construct	First-order construct	Alpha	Echo	Golf
Transition processes	Mission analysis	3.7	3.3	4.3
	Goal specifications	3.5	3.0	4.8
	Strategy formulation and planning	2.3	2.3	3.0
Action processes	System monitoring	1.7	2.0	1.0
	Team monitoring	5.0	2.5	4.0
	Coordination	4.5	4.0	4.0

3.3 Design Feedback

3.3.1 Temporal Structuring and Phase-Based Simulation

The most strongly endorsed design features were those related to temporal structuring, control, and review of operations. The ability to manage COAs by phase (Item 5) received an average rating of 99.5/100, with near-zero variance, indicating strong agreement across participants. Similarly, the ability to simulate operations up to key decision points or by phase (Item 6) earned an average of 97/100. This emphasis on phase-aware planning is directly linked to operational realism. SMEs described how temporal sequencing is essential for modeling cause-and-effect chains in operations to allow shaping and supporting operation or the coordination and massing of effects, such as ensuring artillery has cleared a position before maneuver forces advance or that an enemy is first degraded with fires before a maneuver element engages them directly. Participants elaborated on this need by proposing functionality that enables planners to define tasks and dependencies by phase, simulate plans incrementally, and observe the timing of critical failures (e.g., when supporting units are destroyed or cannot execute follow-on tasks), and provide new interface elements to allow fluid, intuitive review and manipulation of these more granular temporal elements. For example, the SC3 UI currently lacks any timeline interface element to display the sequence of commanded tasks or simulated actions over time for each unit or task organization. This could notionally take the form of a Gantt chart or synchronization matrix, as is used traditionally in MDMP. Furthermore, the mechanisms for viewing and modifying the tasking of units is not currently ergonomic or user-friendly, limiting interaction within a side panel (i.e., not within the map) and to tasking for a single unit at a given time (i.e., no way to see tasking for all units), which limit the ability to ensure proper sequencing and coordination of tasks across units. The current UI does not display this temporal information within the map, showing only the initial and final disposition of forces, with effectively no intermediate data presented (beyond unit movement paths, which also do not have any display of their relationship to time). One participant noted that a system lacking this ability, instead showing only the end state of the simulation, leads to COAs that violate basic doctrinal logic, such as seizing an objective before supporting fires have been emplaced, due to a lack of transparency that such logic has been violated.

To fully support these use cases, SMEs recommended the following:

- Adjusting task assignment timing of multiple units with greater ease of use and precision.
- Toggling and visualizing simulation playback by phase or more granularly.
- Marking decision points and contingency triggers.

- Creating and naming unit-level control measures like echelon-specific phase lines, no-fire zones, and limits of advance, in addition to those inherited from higher echelons.

Fundamentally, mission planning is an exercise in problem solving and design over a multidimensional domain. While these dimensions include abstract factors, like supplies or combat power, the spatiotemporal dimensions are central to mission planning. This means the spatial (map) UI elements are equally as important as the temporal (timeline) UI elements, which are currently lacking in SC3 UI, and that interactions in one dimension must be kept updated and reflected in the other. For example, if a unit's goal location is moved to a more distant point, this should simultaneously be reflected by a corresponding change in the timeline, resulting in a longer scheduled travel time.

Additionally, because the spatiotemporal dimensions are so interlinked and critical to planning, users may likely be inclined to frequently shift between representations (map vs. timeline). So, to improve ergonomics and minimize interface burden, any functionalities that can be integrated within both elements should be implemented. For example, the ability to adjust unit tasking within the map, as well as redundantly being able to adjust that tasking within the timeline, will allow users to focus on whichever of these interface elements is more convenient in the moment, minimizing unnecessary transitions between them, as would be the case if this could only be manipulated in one of the interface elements.

3.3.2 Streamlined Tasking and Unit Control

SMEs rated several features related to tasking and unit manipulation as highly important. The ability to group and manage units at different echelons (Item 3) received a 91.5/100 average rating. This reflects frustration with the current system's requirement to manipulate units individually, especially at company level.

While SC3 Sim does have data files that capture the hierarchy of the 25ID and all subordinate units down to company level, it does not have the capability to dynamically reconfigure the echelon at which the simulated entities are instantiated. Instead, there is a one-to-one relationship between the entities that are visualized, can be tasked, and can be simulated, with no further abstraction or aggregation between simulated entities and how they are displayed or tasked. Given this, SC3 Sim is currently configured to instantiate entities at different echelons, depending on their relevance to the 2/1AD that the SMEs were notionally planning for. Doctrinally, the brigade (e.g., 2/1AD) has command authority to task its subordinate battalions (e.g., 1-1CAV), as well as the companies (e.g., C 1-1) within those battalions, but no lower (e.g., platoon and below), as the tasking of units more

than two echelons below is left to subordinate commanders to dictate. Therefore, because all the companies within 2/1AD could theoretically be separately tasked by the brigade planning staff, they were individually instantiated within SC3 Sim and SC3 UI, while the battalions they comprised were not represented within those systems. Conversely, 1st Brigade Combat Team, 2nd Infantry Division (1/2ID) is an adjacent brigade, not falling under the command of 2/1AD, and thus the SMEs did not have the notional authority to task that unit. So, 1/2ID was represented as a single brigade-level entity rather than being broken out into lower-echelon entities, merely acting as a marker that another unit was present outside 2/1AD's AO.

SC3 Sim and SC3 UI currently do not provide a mechanism for forming ad hoc groupings of entities, tasking them collectively as a group, and then intelligently propagating the group's tasking to the entities within it in a nested and doctrinally appropriate manner. Rather, if a COA is generated by COA-GPT that tasks a unit at a particular echelon, but SC3 Sim represents that unit as its constituent, subordinate units at some lower echelon, the ITS simply assigns that same task to all of those subordinate units, even if that is not suitable for each unit. Furthermore, if SC3 Sim represents a set of entities at a particular echelon, there is no way to change the tasking of all those entities with a single action, and their tasking must instead be changed individually for each entity.

SMEs expressed a strong preference for functionality that allows moving and tasking units at intermediate echelons (e.g., battalions) or other task organizations as cohesive groups. This aligns with real-world command structures and processes, such as the attachment of additional subordinate units to increase the capability of a commanding unit ("plus up") or the formation of a task force from an inorganic combination of units to meet a specific but temporary task and purpose. This flexible reorganization of units is commonly conducted several times within a multiphase operation and therefore is critical to enable within the ITS.

Closely linked to this gap is the ability to edit unit properties and tasking from within the map view (Item 7), rated at 92/100. As stated in the previous subsection, the SC3 UI does not currently have interface elements that allow the ability to view or edit the tasking or properties of more than one unit at a time, with those functionalities limited to a side panel. SMEs indicated that an ability to click directly on units within the map to make these adjustments, rather than navigating dropdowns or nested menus, would allow for faster, more intuitive task assignment and property modification. The drag-and-drop tasking with timing constraints (Item 9, rated 88.5/100) was also seen as important for rapid plan creation, particularly when coordinating movement tasks and synchronizing unit arrivals with mission phases. As stated in the previous subsection, the ability to interact

with these functions within the map and within the timeline (when implemented) would be crucial.

Additional features in this category included the following:

- Copy-paste tasking across units (Item 1, 84.5/100) to reduce repetitive actions for similarly assigned units.
- Zoom toggling between unit-level and operational views (Item 10, 77/100) to support contextual awareness while editing.
- Toggleable resource cost visualization (Item 8, 84.5/100) for sustainment and resourcing insights.

SMEs further recommended integrating search and follow functions (e.g., “Where is B-1-1?” or “Track D-47 during Phase 2”), highlighting how situational awareness suffers when users must scroll or pan manually to track unit behavior. This is particularly difficult in this domain where planners must track the actions and states of many entities over time and space.

3.3.3 Natural Language Editing and COA-GPT Refinement

SMEs placed high value on expanding natural language editing via COA-GPT (Item 4, 86/100). They envisioned using plain language to make specific, targeted changes to COAs, such as “Delay movement of C 1-1 until Phase 3” or “Replace the recon team with a support-by-fire position.” This interaction mode was seen as essential for enabling users to refine COAs at speed without manually rebuilding task sets in the interface. COA-GPT does have some prototype capabilities for using natural language voice interacting and generalized prompting (not directed toward a specific MDMP substep or product), but these were nascent features not utilized during the touchpoint.

Natural language interaction using voice would also be bolstered if it was fused with other modalities of interaction such as drawing, gesture, or even eye gaze. This would allow planners to use the system intuitively and quickly communicate information to the ITS that has both high-level and rich meaning and is grounded to specific locations in space or the interface for improved specificity and context. For example, being able to have additional information or options appear based on gesture or gaze actions and then interacting with those additional elements through voice. These concepts are being explored in our OpsVision effort. Similarly, the abilities to draw on the map and vocalize tactical mission tasks and have that be communicated to the system as unit tasking and presented as standardized mission graphics would allow for fluid design and editing of COAs. This capability is currently being pursued by the 6.1 research team, in partnership with an industry

collaborator, to integrate COA-GPT with a library for fused sketch and natural language interaction.

SMEs also noted limitations in the current COA-GPT responsiveness, especially when attempting edits tied to specific phases or geographies. The recommended expanding traceability (such that users can click on a unit or task and ask, “Why did you do this?”) and providing clear links from each generated action back to the source order or doctrinal justification. This would greatly improve transparency and support doctrinal validation during planning.

3.3.4 Adversarial and Collaborative Planning Support

The ability to wargame against another human adversary (Item 11) was rated at 95.5/100 with almost no variance across SMEs, indicating high levels of priority and agreement. This feature was described as critical for improving the realism, adaptability, and rigor of COA testing, rather than treating the simulation as a static validator of blue plans. SMEs envisioned a dynamic planning environment where users could do the following:

- Control opposing sides of the battlefield.
- Introduce deception, maneuvers, and counters.
- Pause, annotate, and revise COAs during execution.
- Replan in response to enemy decisions.

This would align the ITS functionality more with the doctrinal process of wargaming, a substep of COA analysis within MDMP. In that substep, planners work through a developed COA in a step-by-step evaluation process, where a portion of the planning staff act on behalf of the friendly forces, another portion act on behalf of the enemy forces, and a third contingent act as a neutral arbiter between them. In a turn-based manner, the friendly forces perform an action, the enemy force perform a reaction, the friendly forces perform a counteraction, and the expected outcome of each action are adjudicated (typically in a slow, subjective, and manual process), with this cycle continuing until completion or a decision to redesign is made. This differs from how the ITS is currently designed to function, in which COAs for friendly and enemy forces are generated and fully simulated to their end (enabled by the rapid pace of simulation), with no opportunity for friendly or enemy intervention during the course of the COA. Implementing the functions described by the SMEs, would likely yield the benefits of adversarial planning (a more experienced, creative, and robust opponent and the capability to generate and explore branch and sequel plans throughout the intermediate steps of a COA), while benefiting from the rapid speed, increased objectivity and rigor, and ability to track and assess a wider set of COA branches and sequels. Furthermore, the ITS could

be segmented into friendly and enemy instantiations such that each respective contingent of planners are limited to only the information available to their forces (i.e., fog of war), rather than a mission state that includes a complete ground truth, as is the case with traditional wargaming. This would allow planners to test their COAs in more challenging settings where they have incomplete information about their opponent's composition, disposition, and actions. These features would directly support training and decision rehearsal use cases and better reflect the adversarial and uncertain nature of real-world operations.

For collaborative planning, SMEs emphasized the importance of structured role and area ownership. They proposed that users should be able to assume responsibility for geographic areas (e.g., north sector of the AO), organizational elements (e.g., battalion task force), or functional domains (e.g., fires or logistics). This would enable distributed planning, mirror present-day command post organization, and allow both synchronous and asynchronous workflows. SMEs also stressed that all changes must be traceable by user, with clear visibility into who made what edits and when, to ensure accountability and coordination. These notions of interfaces and processes that support role- and function-specific use, dynamic transition between taskwork and teamwork, and user-specific customized traceability are central to our OpsVision research effort. While we did not manage to implement a capability for collaborative COA editing within the SC3 UI in time for this touchpoint, the choice to leverage a browser-based UI was so that multiple users could easily share and interact a common set of COAs on a local (or remote) network, enabling them to work independently or collaboratively, with user traceability and COA design progression natively captured.

3.3.5 Additional SME Recommendations

In their open-ended comments, SMEs suggested several additional enhancements.

3.3.5.1 Prompt or Checklist to Validate Unit Roles and Tasking

This gap highlights a significant area of research that will be necessary for the success of this system, specifically automated features for diagnostic characterization of various elements of a COA or simulation configuration to detect potential violations or malformations, and either alerting the user or intelligently resolving them.

3.3.5.2 Notification When the Simulation Completes Running

This initial prototype implementation of SC3 Sim did not have methods for checking the status of running tasks and terminating the simulation rollout when they were complete, nor did it have mechanisms for assessing triggers for premature

termination; e.g., if a failure conditions for more than 40% of the main efforts combat power being attritted. This meant that during the touchpoint, SC3 Sim would run all COA simulations for a fixed duration chosen to safely encompass the execution of all possible tasks, which resulted in a significant portion of each simulation's run time being after the point of task completion and effectively wasted time. The speed, and thus efficiency, of the system would be dramatically improved if there were methods to terminate simulation when no further valuable information can be gathered.

3.3.5.3 Naming of All Graphics and Control Measures

This point of feedback highlights a need for UI design that is complete, clear, and ergonomic. Though all mission graphics within SC3 UI did have their associated names displayed, they were not necessarily always visible depending on what portion of the map the user was viewing (i.e., they did not dynamically move to remain in view), they were not presented in a way that made them easily distinguished from other label (i.e., they were not visually distinct from the unit labels), and they did not follow doctrinal standards for the presentation of military graphics. More broadly, this issue emphasizes that a lack of investment in the visual design and function of a UI (often considered a secondary concern) can significantly hinder users and the efficacy of a system.

3.3.5.4 Support for Engineering and Sustainment Functions, Such as Obstacle Placement, Resupply, and Casualty Replacement

A major limitation of the underlying mechanics of SC3 Sim—in this initial prototype version—is that it lacked elements that capture warfighting functions beyond movement and maneuver. For example, it did not include elements critical to several warfighting functions, such as the following:

- Sustainment: modeling the consumption of supplies (e.g., munitions, fuel, end items), tasks and behaviors for resupply of units, etc.
- Engineering: environmental features and effects representing obstacles, task and behaviors for emplacing and removing obstacles, etc.
- Fires: mechanics for indirect fire (via artillery) that requires coordination with a separate ground unit to directly observe the target, etc.
- Intelligence: partial observability of the enemy's composition and disposition, reconnaissance units that can be tasks to resolve enemy observability, etc.

These mechanics, while not as fundamental as movement and maneuver, are an essential next step to ensure there is further value in SC3 Sim for COA analysis, and particularly wargaming.

3.3.5.5 Search Tools for Quickly Locating Units or Assessing Task Status

This point of feedback emphasizes previous themes: 1) that the ability of users to sufficiently monitor and evaluate the state of so many entities over time and space is particularly challenging in this domain of LSCO planning; and 2) that solutions to this problem will require both well-designed, ergonomic interfaces and sophisticated algorithms for diagnostic characterization of the COAs. Design characteristics that support human-machine teaming, fluid interaction, and decision support are just as critical to the success of these technologies as the computational capabilities to rapidly generate and forecast COAs and cannot be relegated to secondary or tertiary considerations.

3.3.5.6 Visualization of When Units Failed or Succeeded in Executing Their Missions, Aligned to Time and Space

This feedback also points to multiple previously stated themes: 1) that users will need algorithmic assistance to assess the state of a COA, triage critical events, and alert them to critical decision factors with clear traceability; and 2) that all critical data visualization and interaction elements must be tied to both the spatial presentation (map) and temporal presentation (timeline), where possible. The process of iterative analysis and refinement, which is fundamental to the design of effective COAs, requires that planners can make modifications to COAs and easily understand the effects on outcomes, so as to efficiently explore the planning space. Visualization and decision support features of this technology must aid in identifying and transparently presenting these effects for them to be beneficial to planning. COAs that are inexplicable, no matter how rapidly produced or analyzed, will not support planning.

Participants also identified minor errors in COA behavior, such as unintended switching of maneuver types (e.g., a phased envelopment being altered to a phased frontal attack). These issues, while technical, underscore the importance of transparency and validation tools that allow users to audit changes and maintain doctrinal fidelity.

3.3.6 Priority Simulation Mechanics Design Goals

In addition to the SME feedback captured during AAR, the collaborative design discussions conducted during the touchpoint helped to shape a set of priority areas for future development and implementation approaches, focused primarily on updates to the simulation mechanics.

3.3.6.1 Representation of Units in Terms of Constituent Classes with More Granular Tracking of Their Properties and State

SC3 Sim currently does not represent units in terms of their composition at the level of entities or capabilities. A priority goal in the next version of SC3 Sim is to configure units in terms of asset classes that represent the individual entities that comprise them. These constituent assets would be designed such that they sufficiently define general and ubiquitous building blocks for composing any unit type, delineate entities with distinct capabilities and planning considerations, and without fully enumerating all possible variants of every piece of equipment. Each of these asset classes would also have associated loadouts and consumption rates of Class I (rations), Class III (petroleum), and Class VIII (medical materials) as a function of time, distance traveled, and/or engagement state (engaging or engaged). Furthermore, each of these asset classes would be defined by their payload in terms of armament classes, which again would be designed to capture their distinct implications for lethality and sustainment, while acting as general groups of weapon systems. Each of these armament classes would be defined in terms of their maximum effective range, rate of munition depletion (rounds expended per hour), and rate of attrition of enemy asset classes (entities destroyed per hour) when engaging threats during combat.

For example, a mechanized infantry company might be defined in terms of its composition of asset classes: 14 armored fighting vehicles (e.g., M2A3), 1 armored personnel carrier (e.g., M113A3), 1 prime mover (e.g., M1078A1), 2 utility vehicles (e.g., M1165A1), and 78 dismounted Soldiers. Furthermore, each armored fighting vehicle might be defined in terms of its composition of armament classes: 2,200 rounds of machine gun (e.g., M240C), 900 rounds of autocannon (e.g., M242), and 2 missiles (e.g., BGM-71 TOW).

To support planning considerations related to the sustainment warfighting function, such as resource consumption and resupply logistics, it is critical that we track the state of units with more specificity and precision than the aggregate values of strength (percentage) or combat power (unitless). By expanding the representation of units to capture specific asset classes that comprise them and their corresponding armament classes, we can more directly track the losses (i.e., casualties and destroyed vehicles), and consumables (munitions, fuel, rations, etc.). Figure 12 provides a visual depiction of how this future concept (Version 2) would differ from the current implementation (Version 1).

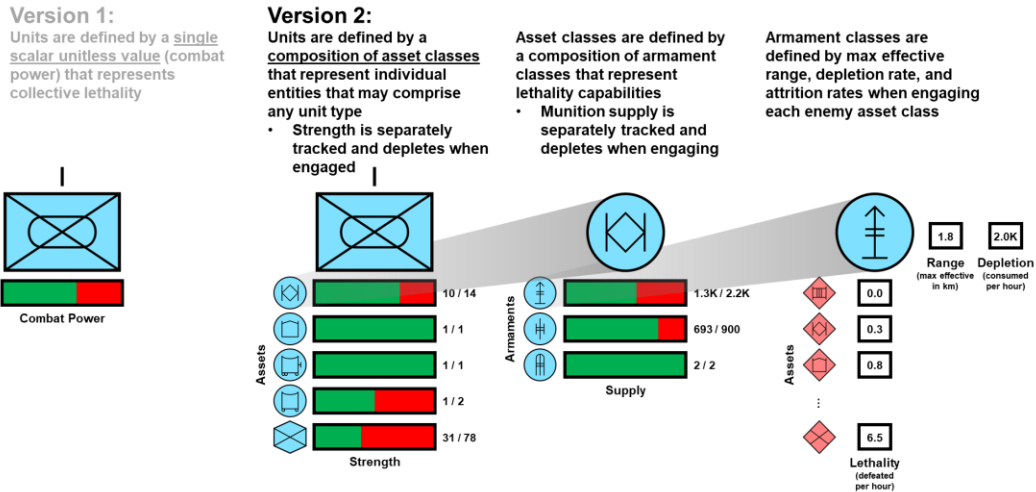


Figure 12. Future unit configuration concept.

3.3.6.2 Representations of Multiple Armament Capabilities and Maximum Effective Range within Units

In conjunction with the improvement just described, wherein units would be defined with more granularity of their constituent assets and armaments, a priority would be to incorporate the varied maximum effective ranges of those armaments into a more granular representation of the unit's combat area. Rather than a singular area of effect for a unit to engage targets, there would be a series of concentric areas that correspond to each of the armaments contained within a unit. The effects of these various armaments to attrit assets within the unit it is targeting would only be enacted when each of these areas respectively begin to overlap with that target unit. For example, attrition of a target unit by cannon fire would begin when the combat area with size equal to the maximum effective range of a cannon armament overlaps with the target unit's formation area, while attrition by small-arms fire would not begin until the distance between them is reduced and the combat area with size equal to the maximum effective range of small-arms armament overlaps with the target unit's formation area. Figure 13 provides a visual depiction of how this future concept (Version 2) would differ from the current implementation (Version 1).

3.3.6.3 Resource Consumption, Simplified Logistics, and Resupply Mechanics

To support analysis and wargaming needs of the sustainment warfighting function, the critical first step is to track the consumption of resources over the course of a simulated COA, which SC3 Sim currently does not compute. The first implementation of this capability would leverage the concept for expanded unit configuration described above. This would define the composition of assets and armaments for all units to capture both the initial loadout and the rate of depletion

for most class of supplies: Class I (rations), Class III (fuel), Class V (munitions), Class VII (end items), and Class VIII (medical materials). These rates of depletion would be dynamically modulated by the state of each unit in terms of time elapsed, movement, and combat engagement. Simply computing these rates of depletion would help planners project likely points of needed resupply or failure of a COA. Building upon this mechanism, we would then implement a simplified approach for resupply. This would involve sustainment units moving to the units they are supplying or vice versa, until their formation areas intersect. The resupply itself (i.e., the increment rate of supplies) could be instantaneous, at a constant rate, or scale with the amount of overlap of the two units' formation areas, depending on what is representative of actual resupply timelines. Furthermore, this resupply mechanic could be passively performed by the relevant units, based on a fixed lower-level policy, or could be a more active planning feature that would need to be specified as tasks within the COA that are performed by specific units at specific times.

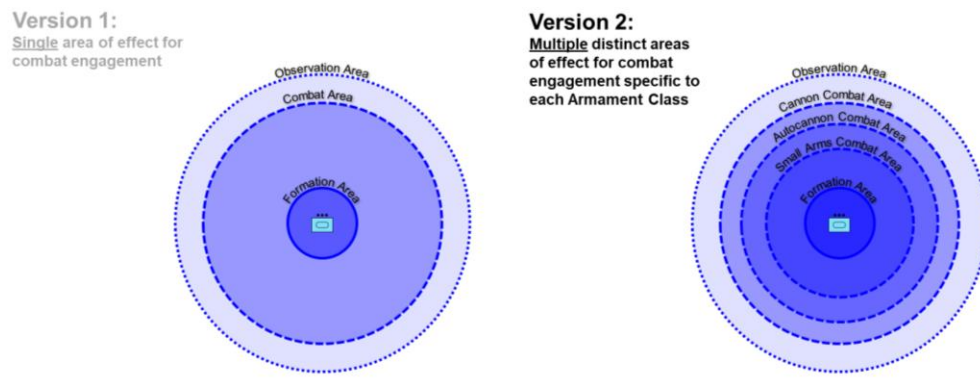


Figure 13. Future combat area concept.

3.3.6.4 Attrition Rates that Vary as a Function of Proximity between Engaged Units

SC3 Sim currently does not modulate the effects of combat based on the proximity of the two engaged entities once overlap between the shooter's combat area and the target's formation area has occurred. However, these circular regions are abstractions representing the areas that individual entities within the unit collectively occupy, can directly observe, or can effectively engage targets. Therefore, we would not expect that the full effects of combat would be applied when only the most distal element of one unit is minimally within range of the most distal element of another unit. To account for this, we will prioritize implementing a modulation factor on the base attrition rate and munition depletion rates that scale with the amount of overlap of these regions. This simple mechanism will capture

the effect of lethality being progressively applied to a target unit as more of their formation falls within an attacking unit's maximum effective range. The use of circular regions allows for computational simplicity, as a closed-form solution exists for calculating the area of overlap between two circles, given their radii and the distance between them. Figure 14 provides a visual depiction of how this future concept (Version 2) would differ from the current implementation (Version 1).

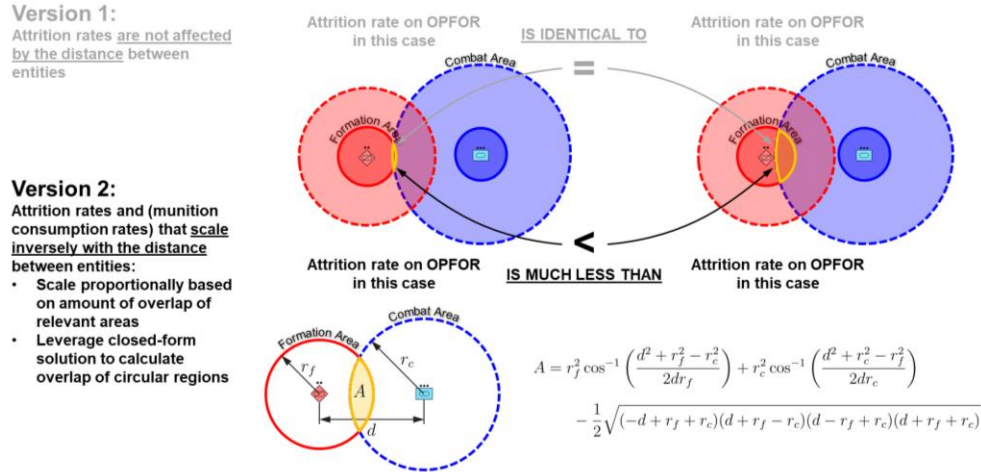


Figure 14. Future range effects concept.

3.3.6.5 Indirect Fire Mechanics that Are Dependent on Observation by Separate Units

The current implementation of SC3 Sim treats all entities as direct-fire units and does not reflect capabilities for indirect fire, such as field artillery, which have larger maximum effective ranges at which they can engage targets while relying on some other friendly unit to provide observation of the target being engaged. To more accurately represent such a capability, we will implement a mechanic wherein enemy units can only be engaged by field artillery if there is simultaneous overlap of the artillery unit's combat area, a separate friendly unit's observation area, and the enemy unit's formation area. If multiple instances of these conditions exist simultaneously, then the lethality effects of the artillery unit will be evenly split between each instance, as is currently implemented for direct fire engagements. More sophisticated mechanisms for target allocation management can replace this simplified approach as needed to support planning the fires warfighting function. Figure 15 provides a visual depiction of how this future concept (Version 2) would differ from the current implementation (Version 1).

Version 1:
No mechanic for indirect fire (IDF)

Version 2:
Artillery units have a much larger Combat Area than Observation Area and require a separate unit to observe an enemy in order to engage that enemy with indirect fire.

Areas must overlap to engage the enemy with indirect fire:

- Artillery Combat Area
- Observer Observation Area
- Enemy Formation Area

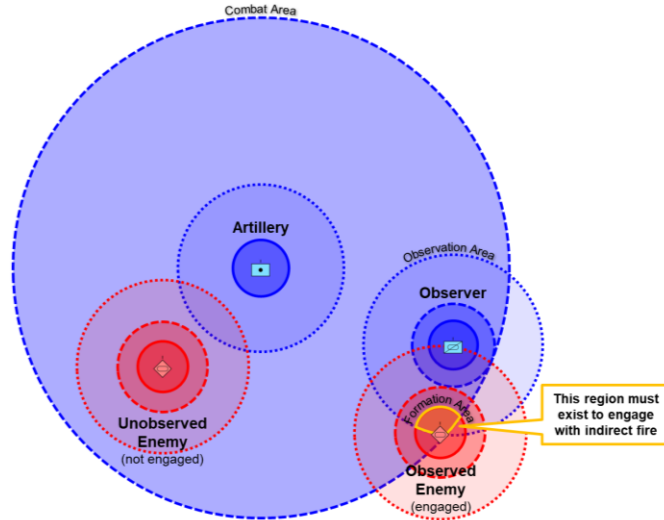


Figure 15. Future indirect fires concept.

3.3.6.6 Terrain Occlusion of Observation and Combat

The current implementation of SC3 Sim does not represent the occlusion of line of sight (LOS) by terrain features. The simulation is effectively capturing the COA outcome as if the terrain were completely flat. This simplifying assumption omits an important planning consideration and modifier of the dynamics of combat engagements, in that terrain occlusion can dramatically impact outcomes. To account for this effect, we will prioritize implementing a mechanic to check for unbroken LOS between opposing units and then prevent the effects of observation and combat in the case that terrain occludes them. This can be straightforwardly achieved by leveraging various ray tracing libraries. The most accurate but computationally costly approach would be to use a fully 3D terrain mesh and perform LOS checks on each simulation tick between opposing units whose observation conditions are already met (i.e., relevant formation and observation areas collide). A less accurate but possibly less computationally complex (and thus faster) approach would be to create a 2.5-dimensional map, where the terrain contours are treated as flat, polygonal plateaus and LOS is considered obstructed if any such plateaus of higher elevation than the units in question are between them. Figure 16 provides a visual depiction of how this future concept (Version 2) would differ from the current implementation (Version 1).

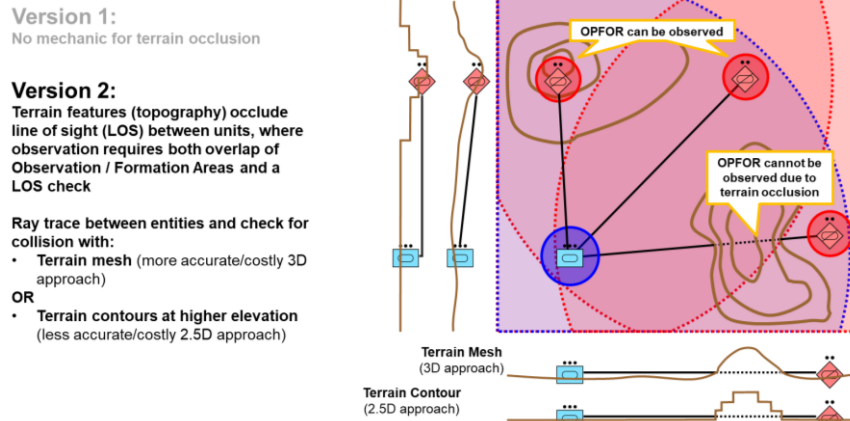


Figure 16. Future terrain occlusion concept.

3.3.6.7 Terrain Classification into Many Types with Varied and Unit-Specific Impacts on Navigation, Mobility, and Observability

The current terrain representation within SC3 Sim is only a binary map, categorizing regions of the AOI as either navigable or nonnavigable terrain. These regions were produced by a terrain classification algorithm using a composite of imagery and depth scan data for the AOI of Operation Tropic Tortoise. This results in a highly simplified terrain effect on simulation outcomes and limited benefit of terrain considerations while planning. Future iterations of SC3 Sim will implement a more complex terrain model, leveraging the multiclass labels already produced by the classification algorithm. This more expansive set of terrain regions would apply disparate effects on navigation cost, movement speed, and observation range, which would be unit-dependent. Rather than units moving only at a constant (though unit-specific) speed, they would have their speed fluctuate based on modifiers for each terrain class (which would also be unit-specific). This provides a richer result for planners to consider, particularly the movement and maneuver warfighting function. Figures 17 and 18 provide visual depictions of how this future concept (Version 2) would differ from the current implementation (Version 1).

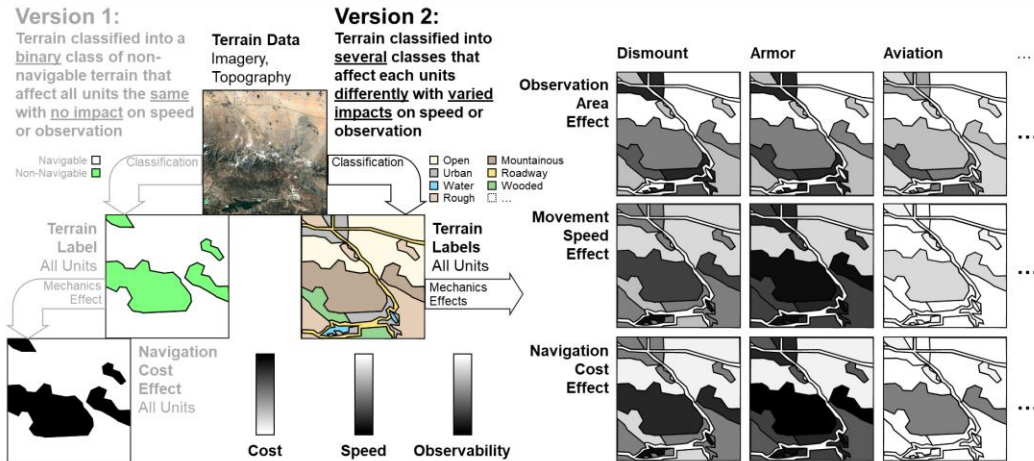


Figure 17. Future terrain classification concept.

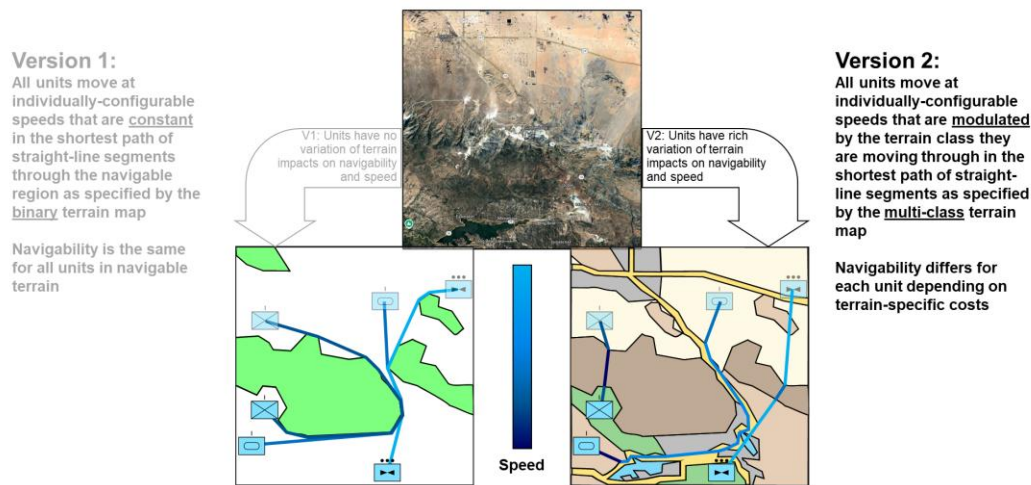


Figure 18. Future movement effects concept.

3.4 COA Design Quality Metrics

We analyzed COAs via simulation quality metrics collected across the three planning sessions to assess early indicators of planning effectiveness, adaptability, and engagement patterns with the ITS and, specifically, the simulation. Metrics included combat power trajectories, destroyed unit ratios, and Blue/Red combat ratios across three participants (Alpha, Echo, and Golf). These provide insight into tactical decisions, engagement timing, and user strategy. At a high level, the data suggest distinct user strategies:

- Alpha demonstrated more aggressive tactics, with sharp shifts in force outcomes often occurring later in the sessions.
- Echo showed stable patterns with intermittent engagement spikes, suggesting a flexible but controlled planning style.

- Golf consistently achieved favorable force ratios with minimal friendly losses, indicating a methodical and effective approach.

These COA quality metrics are presented below over the course of single planning sessions for each of the SMEs. Each marker in the following figures represents a single simulation execution by a SME, which is typically completed within approximately 10 s. Figures are organized by SME participants—one trace for each—and show enemy (Red) and friendly (Blue) force metrics over time, including destroyed unit ratios and combat power trajectories. These visuals reflect how participants engaged with the system to refine their COAs and provide a lens into planning style and tactical effectiveness.

3.4.1 Session 1

Figure 19 illustrates the fraction of destroyed units for the COAs over time for each participant. Red lines represent the destroyed unit ratio for Red forces, and blue lines represent the destroyed unit ratio for Blue forces. Each subplot corresponds to one participant: Alpha, Echo, and Golf. Echo maintains a consistent destroyed unit ratio for both Blue and Red forces across the session, indicating balanced engagements. Golf demonstrates a steadily increasing destroyed unit ratio for Red forces toward the end, while the Blue ratio remains low and slightly declines, suggesting effective offensive operations with minimal friendly losses. In contrast, Alpha exhibits more pronounced fluctuations, particularly a sharp spike in the Red destroyed ratio during the final COA, reflecting a late-session push that resulted in significant enemy losses. These patterns may indicate varying tactical approaches, with Golf emphasizing efficiency, Echo maintaining stability, and Alpha adopting a more aggressive strategy toward the end.

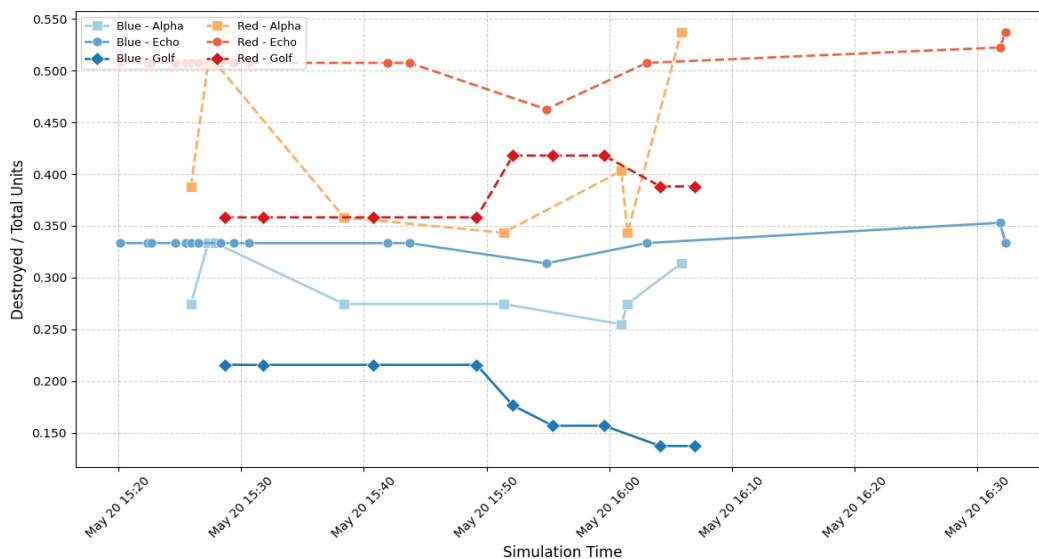


Figure 19. Session 1 destroyed unit fraction.

Figure 20 presents the final combat power at the end of the simulated COA. The top row shows Blue force combat power, the middle row shows Red force combat power, and the bottom row shows the Blue/Red combat power ratio. Echo experiences a sharp drop in both Blue and Red combat power during the final simulation window, pointing to a significant engagement event. The absence of a corresponding spike in destroyed unit ratio suggests either a controlled withdrawal or a short duration. In contrast, Golf maintains steady combat power with moderate attrition, and Alpha shows consistent force levels, reflecting a more measured and continuous operational strategy.

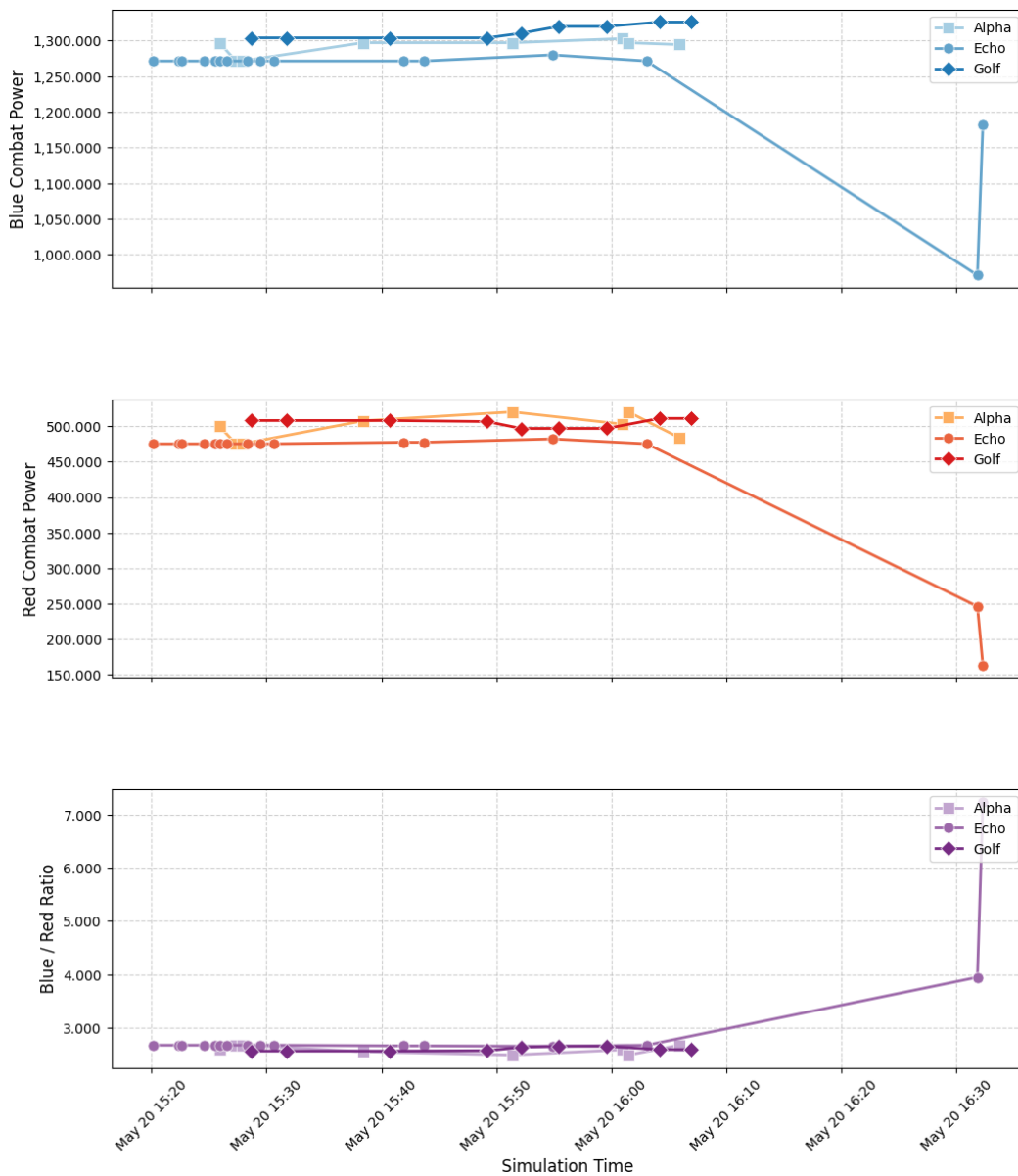


Figure 20. Session 1 final combat power.

3.4.2 Session 2

Figure 21 displays the destroyed unit fraction for planning session 2 for the three participants. Across all three, Blue forces consistently maintain lower destroyed unit ratios than Red forces, indicating superior survivability and effective force preservation. Golf shows a steadily increasing Red destroyed ratio alongside a continuous decline in Blue losses, suggesting escalating operational success with minimal own-force attrition. Alpha also demonstrates rising Red losses over time while keeping Blue losses relatively low and stable, particularly in the latter half of the session. Echo, while maintaining consistently low Blue losses, exhibits greater variability in Red destroyed ratios, with a noticeable upward trend during the final simulation windows. This pattern may reflect increased engagement intensity or pressure during the closing COAs.

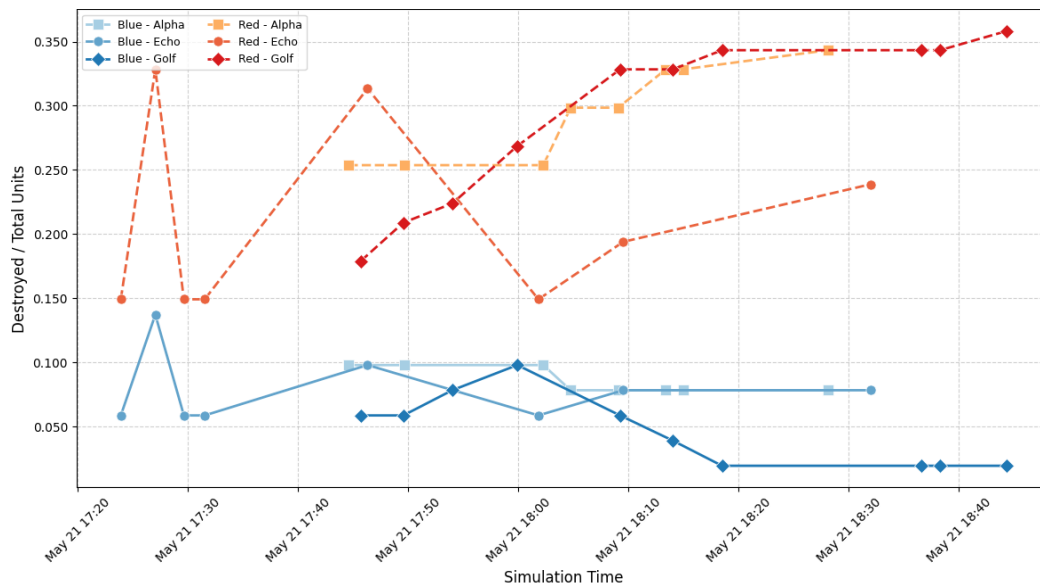


Figure 21. Session 2 destroyed unit fraction.

Figure 22 presents final combat power for planning session 2. Golf shows a steady decline in Red combat power along with a gradual increase in Blue combat power during the middle and late phases. This trend results in a rising Blue-to-Red ratio, indicating a growing tactical advantage. Echo exhibits more fluctuation in both Blue and Red combat power without a consistent directional trend, suggesting variable engagement intensity. Alpha demonstrates relatively stable force levels overall, but with a noticeable increase in the Blue-to-Red combat ratio during the latter stages, reflecting an improving operational position over time.

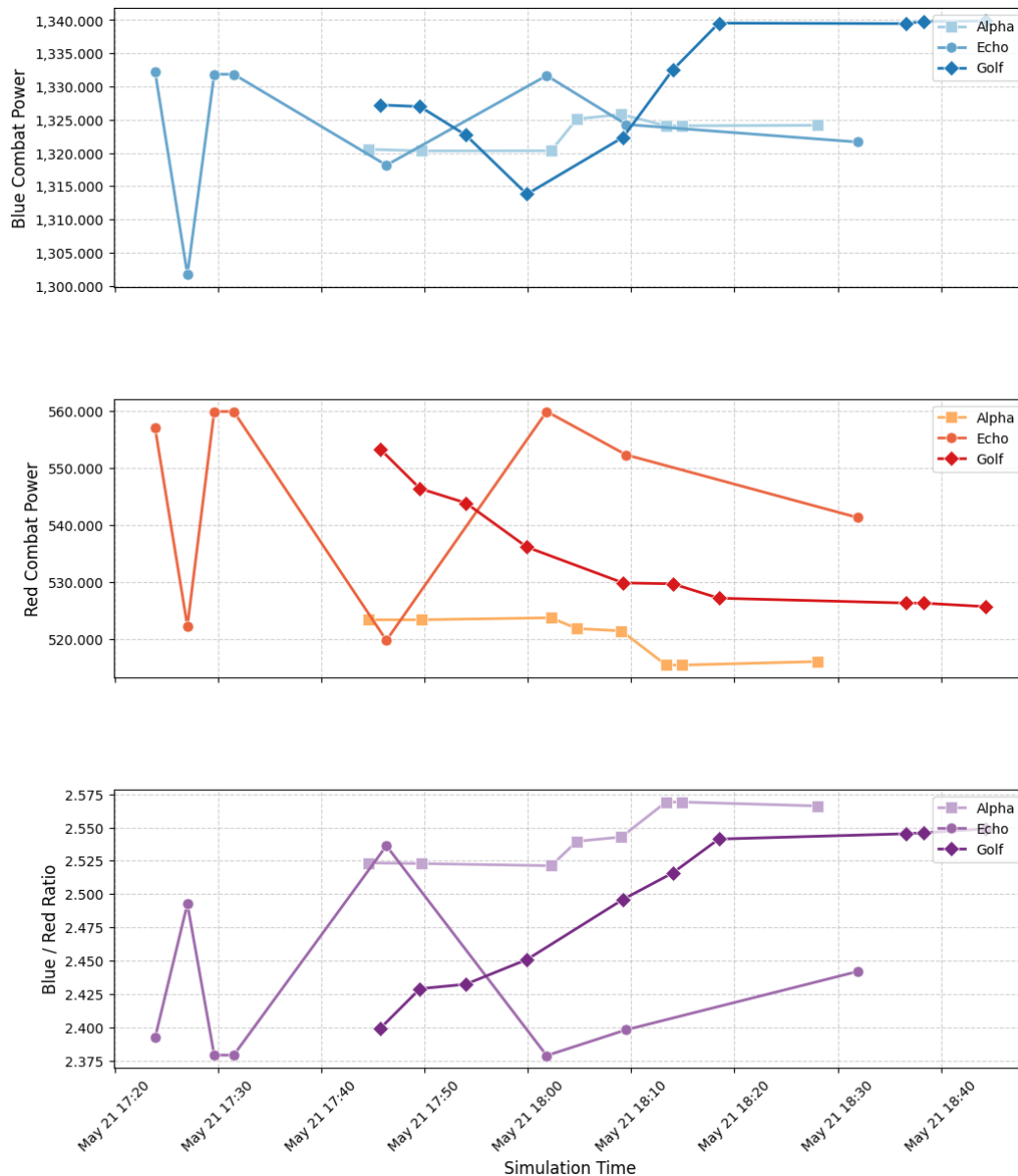


Figure 22. Session 2 final combat power.

3.4.3 Session 3

Figure 23 shows the destroyed unit ratio over time for each participant. Echo maintains a low destroyed unit ratio for both Blue and Red forces for most of the session, followed by a sharp increase near the final timepoint, indicating a sudden and intense engagement. Golf displays a decrease in both Blue and Red ratios, though only two data points are available, limiting interpretation. Alpha shows a sharp rise in Red force attrition early in the session, followed by minimal change, suggesting early engagement and limited combat activity afterward.

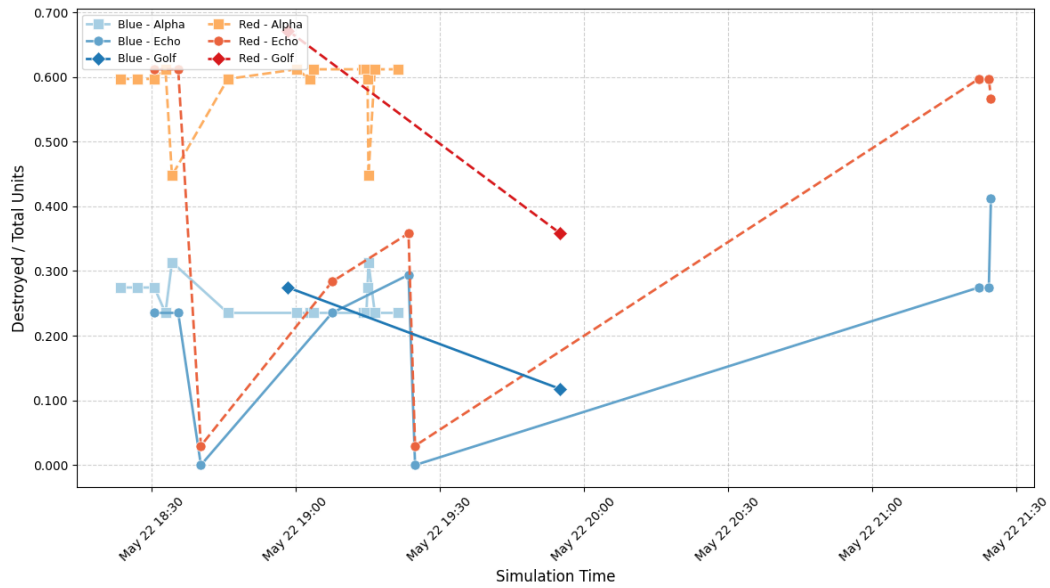


Figure 23. Session 3 destroyed unit fraction.

Figure 24 presents combat power over time for each participant. Echo shows a late-session drop in both Blue and Red combat power, consistent with a final high-intensity engagement. Golf exhibits a steady increase in Blue and Red power through the session, but data are limited to two time points, constraining interpretation. Alpha maintains relatively stable Blue and Red combat power with minor fluctuations. In the bottom panel, the Blue/Red combat ratio rises sharply for Echo toward the end, indicating a relative advantage gained during the final engagement window.

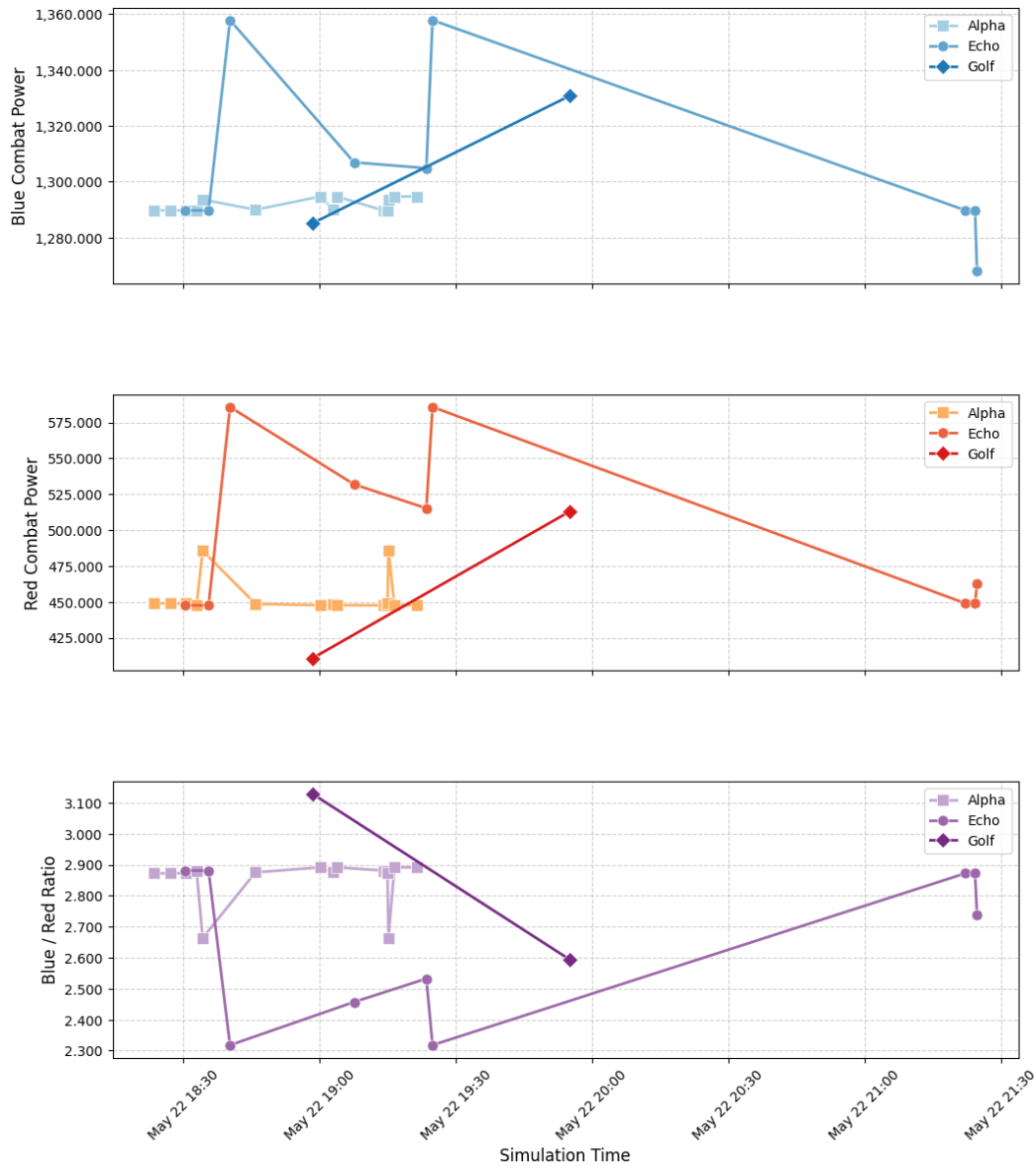


Figure 24. Session 3 final combat power.

3.5 Interface Interaction Metrics

To complement the COA quality metrics analyzed above, we examined how users interacted with the SC3 UI across three days (May 20–22), focusing on their navigation styles and interface focus. These interaction metrics provide insight into users' cognitive strategies, planning approaches, and evolving tactical priorities as they engage with the system during COA development and refinement.

SC3 UI interaction characteristics were analyzed with respect to four core areas of interest (AOI) within the interface, shown in Figure 25: A) COA Creation page, B) Edit Positions page, C) COA Analysis page, and D) Operation Overview page.

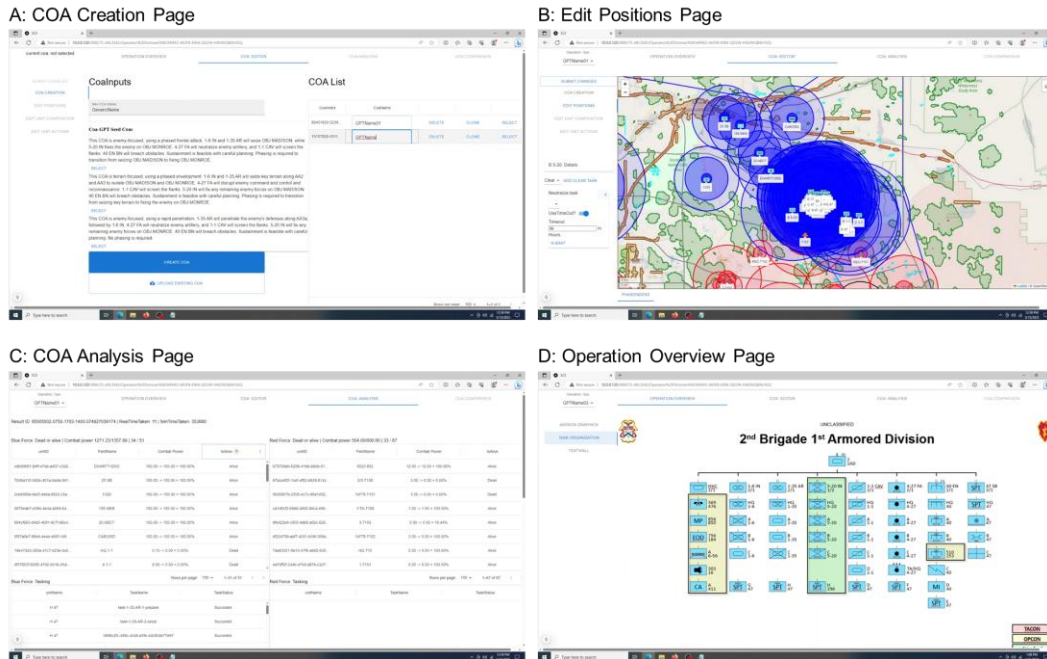


Figure 25. UI core areas of interest.

Each figure in the following section summarizes participants' interaction with the SC3 UI across four core AOIs. Figures are organized by participant and day, illustrating how users navigated and allocated attention across the interface. These patterns provide insight into planning style, interface fluency, and system exploration. At a high level, the data suggest distinct interaction strategies:

- Alpha demonstrated a methodical interaction style, balancing time between Creation and Edit Positions with growing use of analysis tools over time.
- Echo engaged broadly across all AOIs with frequent transitions, reflecting a flexible and exploratory planning strategy.
- Golf remained almost exclusively in the Edit Positions AOI, indicating a focused, execution-oriented approach with minimal interface exploration.

3.5.1 Session 1

Figure 26 presents the proportion of time each participant spent in the four core AOIs:

- Alpha spent approximately 86% of the time in Edit Positions (B) and 13% in Creation (A), with limited engagement in Analysis (C) or Overview (D). This indicates a strong emphasis on tactical layout with minimal scenario exploration.
- Echo displayed a more balanced profile, with approximately 68% in Edit Positions, approximately 27% in Creation, and some presence in Analysis

and Overview. This pattern reflects broader engagement and more exploratory use of the interface.

- Golf spent approximately 97% of the time in Edit Positions with nearly no transitions to other AOIs. This suggests a highly task-focused and possibly rigid interaction style.

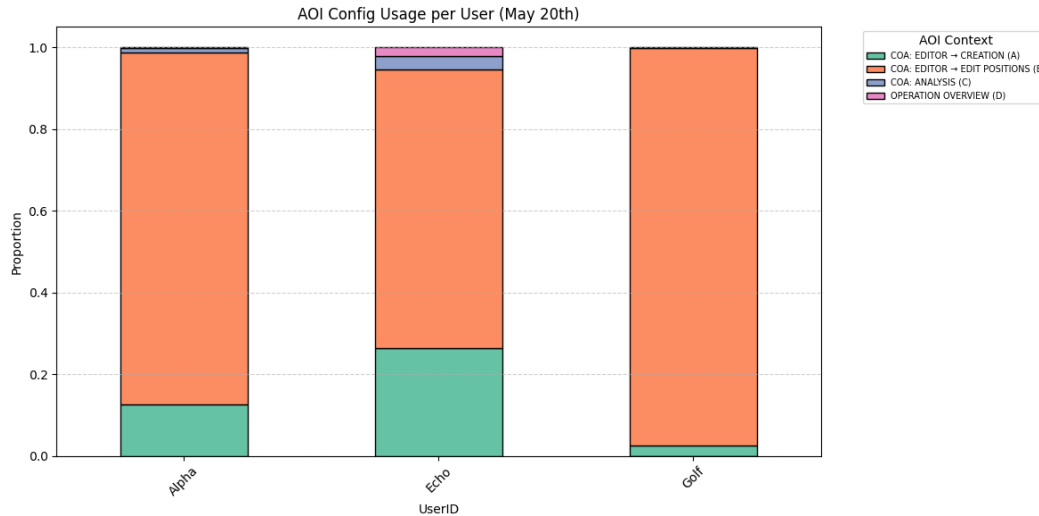


Figure 26. Session 1 AOI usage distribution.

To further characterize user interaction styles, we examined how participants navigated between interface modules (AOIs). Transition probabilities provide insight into interface fluency, decision cycles, and planning behaviors. Figures 27–29 show the transition matrix for each participant in planning session 1. Self-transitions (along the diagonal) indicate sustained focus in a single AOI. Off-diagonal transitions reflect movement across functional modules, typically associated with planning, execution, or situational reassessment. Golf did not navigate to Analysis (C) AOI during this planning session.

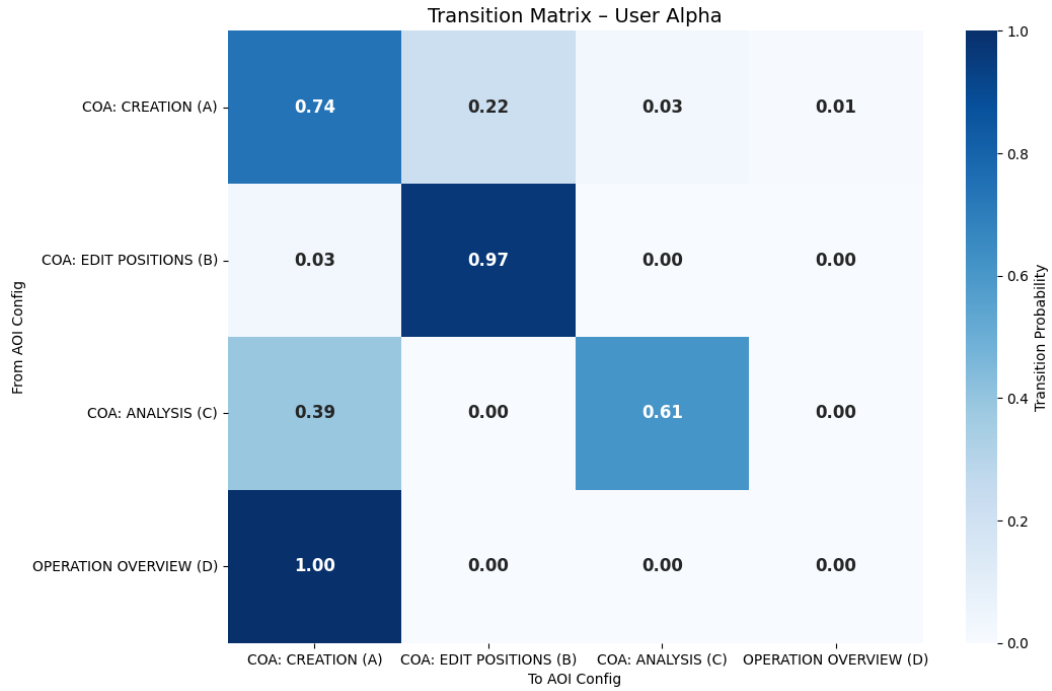


Figure 27. Session 1 transition matrix for Alpha.

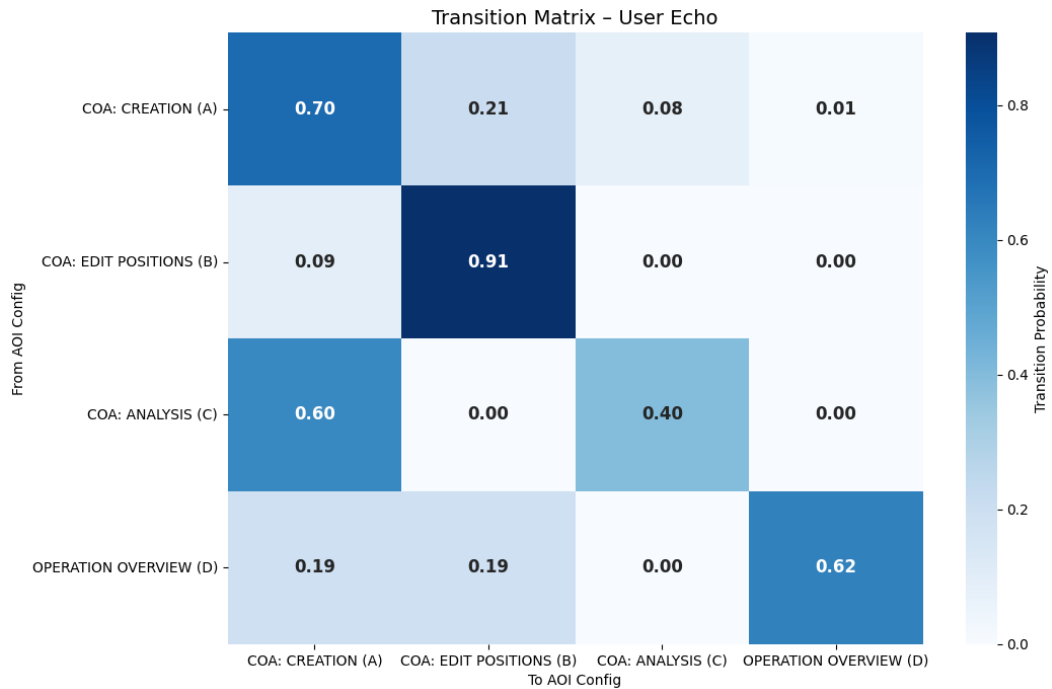


Figure 28. Session 1 transition matrix for Echo.

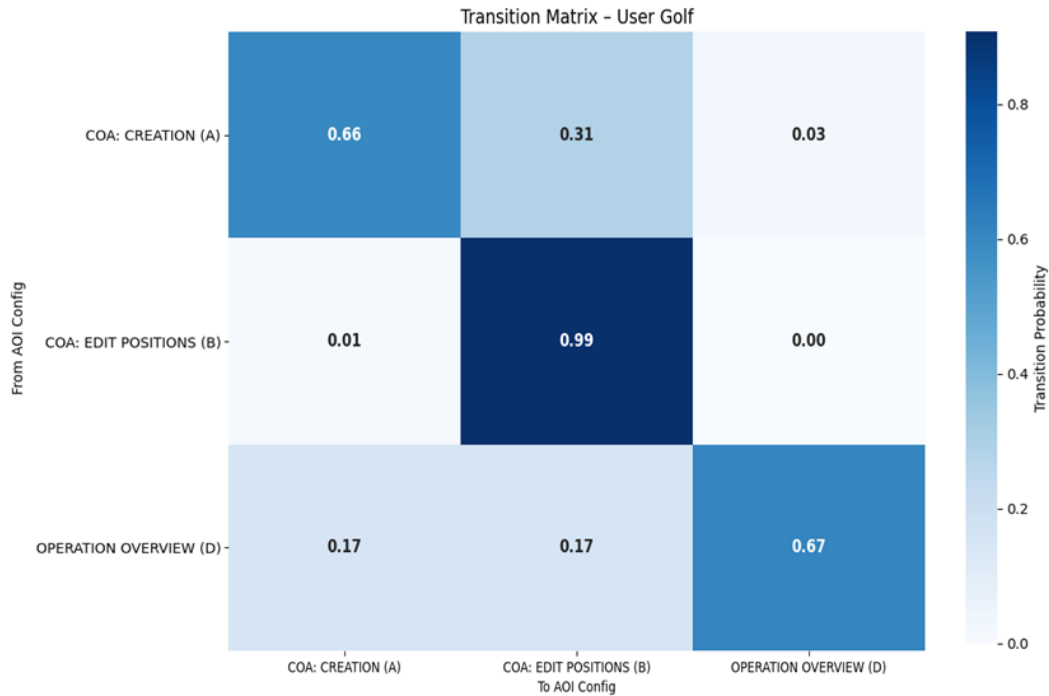


Figure 29. Session 1 transition matrix for Golf.

In planning session 1, participants demonstrated distinct interaction behaviors through their AOI transition patterns. Alpha showed a balanced and iterative workflow with regular switching between Creation and Edit Positions, and occasional transitions into Analysis, suggesting a methodical approach that combined planning with execution and periodic assessment. Echo exhibited the most dynamic transition behavior, frequently navigating between all four AOIs. This indicates an exploratory and adaptive strategy, likely involving scenario testing and reconfiguration cycles. In contrast, Golf maintained a stable pattern, with nearly exclusive use of Edit Positions and minimal cross-AOI movement. This concentrated interaction pattern suggests a narrow focus on execution activities, with limited engagement in planning or analysis modules. Together, these transition matrices reveal early divergence in cognitive engagement strategies, ranging from exploratory and adaptive to highly task-focused and procedural.

3.5.2 Session 2

Figure 30 presents the proportion of time each participant spent in each of the four interface configurations in planning session 2:

- Alpha maintained a similar pattern as Day 1, spending approximately 86% of the session in Edit Positions (B) and 12% in Creation (A). There was a slight increase in transitions, suggesting growing engagement with scenario evaluation or modification.

- Echo continued to demonstrate a more diversified interaction style, engaging not only with Edit Positions (~85%) and Creation, but also showing presence in Analysis (C) and Overview (D). This reflects a deliberate exploration of the interface and suggests dynamic planning and situational monitoring.
- Golf remained highly focused on Edit Positions (~92%), with only limited engagement in Creation (A) and negligible time spent in other AOIs, indicating sustained task execution with little deviation.

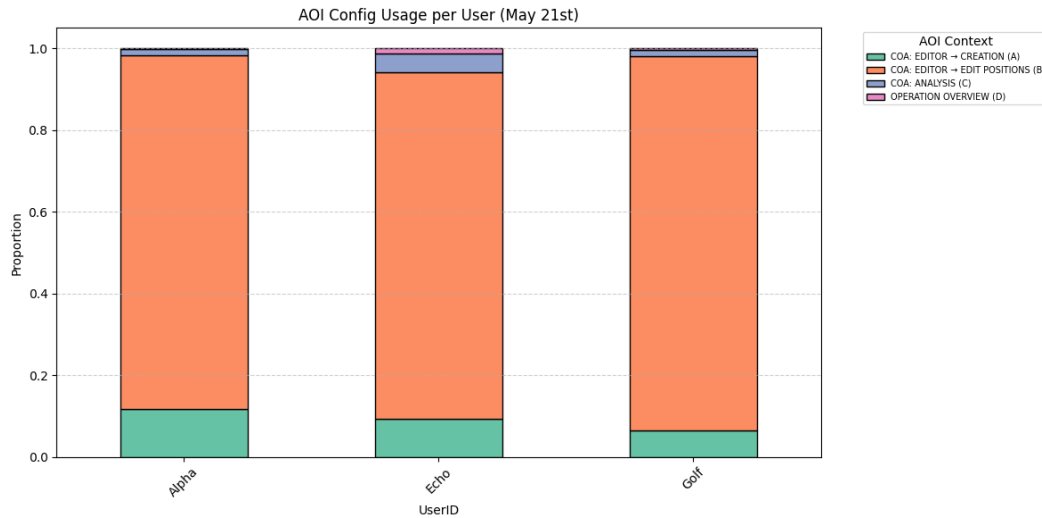


Figure 30. Session 2 AOI usage distribution.

As with planning session 1, we examined transition matrices to analyze the flow of interaction across AOIs. Figures 31–33 present the transition probabilities for one participant.

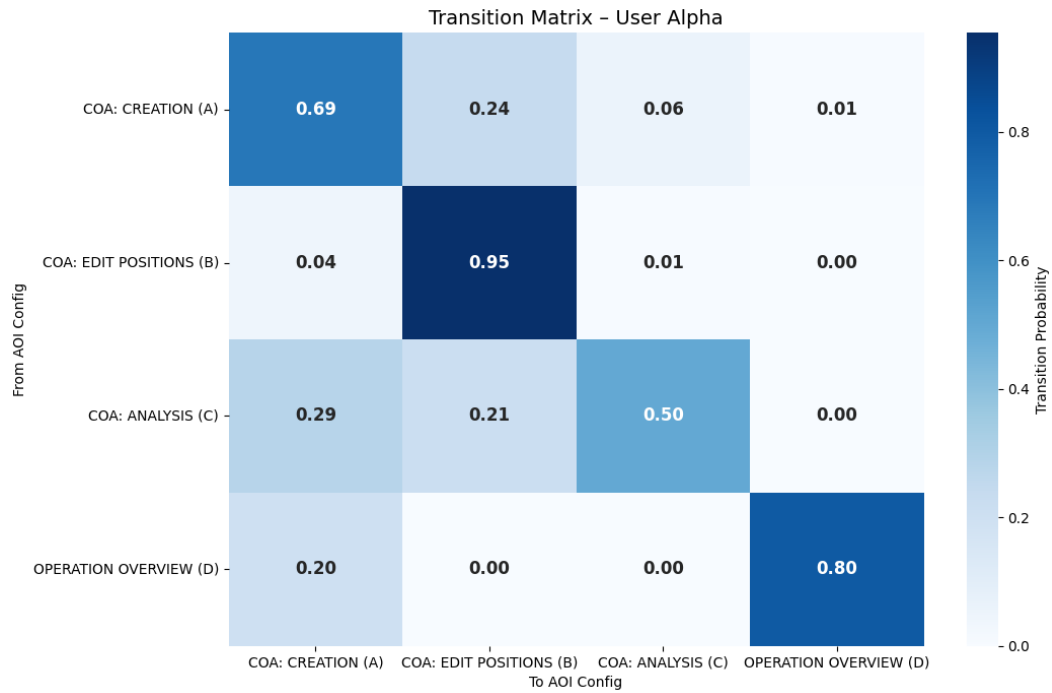


Figure 31. Session 2 transition matrix for Alpha.

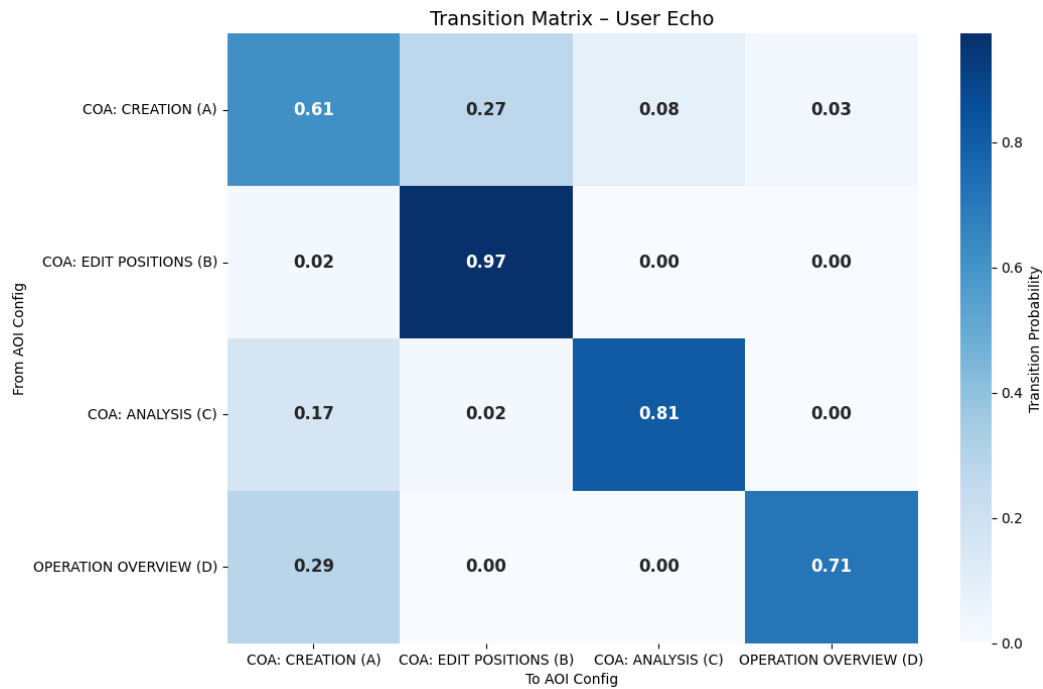


Figure 32. Session 2 transition matrix for Echo.

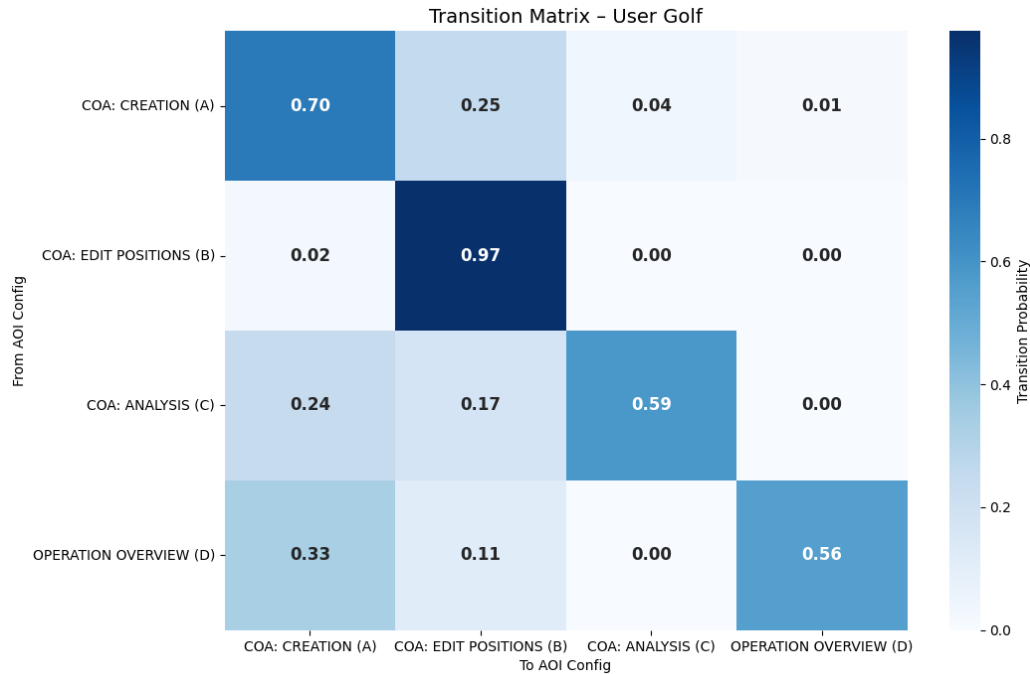


Figure 33. Session 2 transition matrix for Golf.

In planning session 2, participants maintained distinct interaction strategies with subtle yet meaningful shifts compared with the previous day. Alpha showed a modest increase in transitions into Analysis and Overview compared with session 1, suggesting a growing emphasis on evaluating outcomes and refining plans. Echo maintained the most dynamic and exploratory interaction profile, characterized by frequent transitions among all AOIs; including Analysis, Overview, and Creation; indicative of a flexible and adaptive planning process. Golf remained heavily focused on Edit Positions, with 97% self-transitions, but demonstrated a meaningful increase in transitions involving Creation, Analysis, and Overview compared with session 1. This shift suggests a modest broadening of engagement across interface modules; potentially reflecting a move from strictly execution-focused behavior toward limited scenario review or refinement. Collectively, these patterns reinforce the divergence in user strategies; ranging from adaptive exploration and iterative refinement to focused, execution-oriented interaction.

3.5.3 Session 3

Figure 34 presents the distribution of time spent in each interface AOI in planning session 3:

- Alpha shifted toward more use of Creation (A) (~31%) and slightly less time in Edit Positions (B) (~66%), with minimal but present engagement in

Analysis (C) and Overview (D). This suggests an evolving focus toward plan refinement.

- Echo spent the majority of the session in Edit Positions (B) (~87%) with reduced interaction in other AOIs (~12% in Creation, negligible elsewhere). Compared with prior days, this reflects a narrowing of focus.
- Golf maintained strong use of Edit Positions (~75%), with increased time in Creation (A) (~24%). Although Analysis and Overview were still largely unutilized, the engagement with Creation suggests greater involvement in scenario construction or modification.

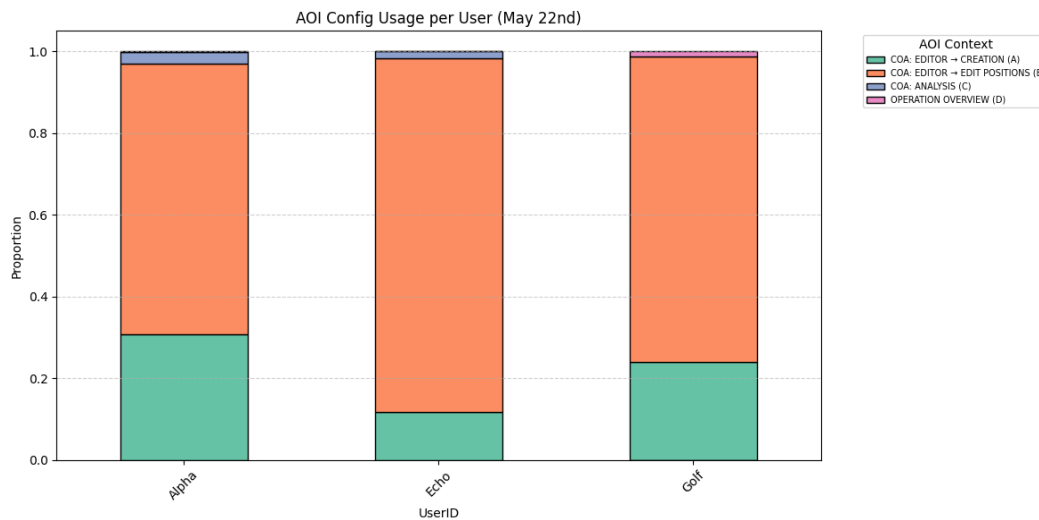


Figure 34. Session 3 AOI usage distribution.

As with planning sessions 1 and 2, we examined transition matrices to analyze the flow of interaction across AOIs. Figures 35–37 present the transition probabilities for one participant. Echo did not navigate to Overview (D) AOI and Golf did not navigate to Analysis (C) AOI during this planning session.

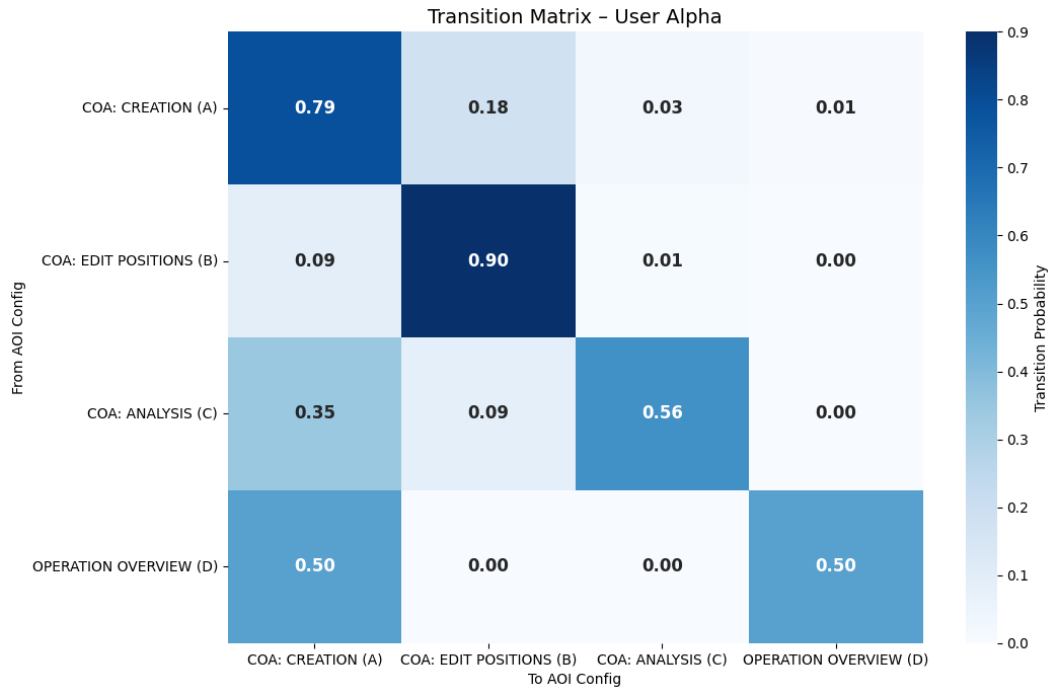


Figure 35. Session 3 transition matrix for Alpha.

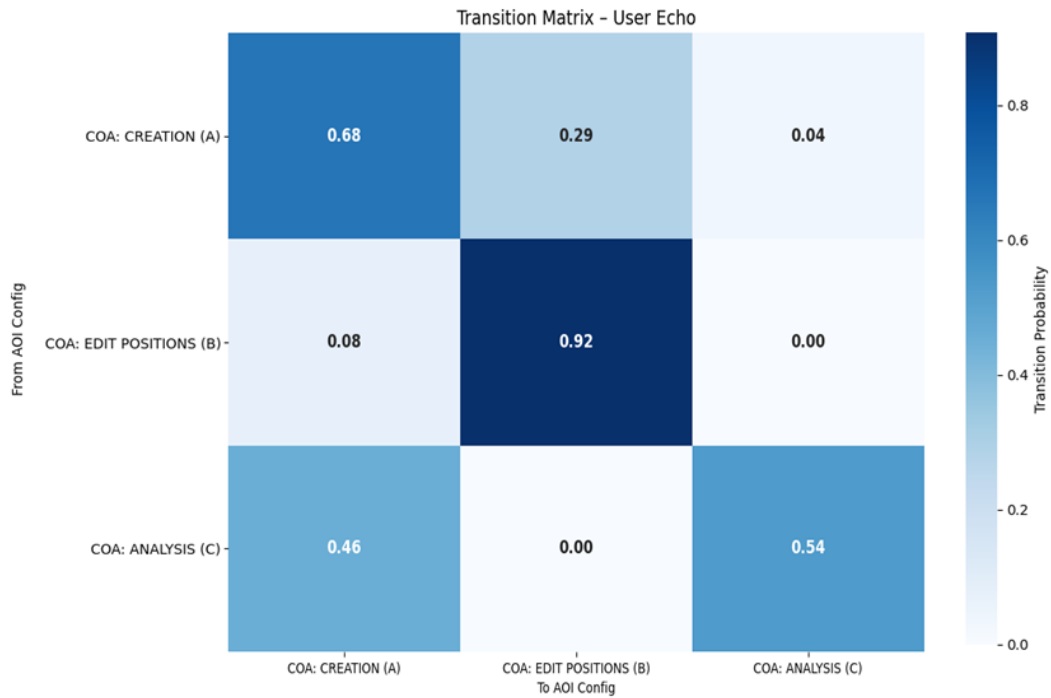


Figure 36. Session 3 transition matrix for Echo.

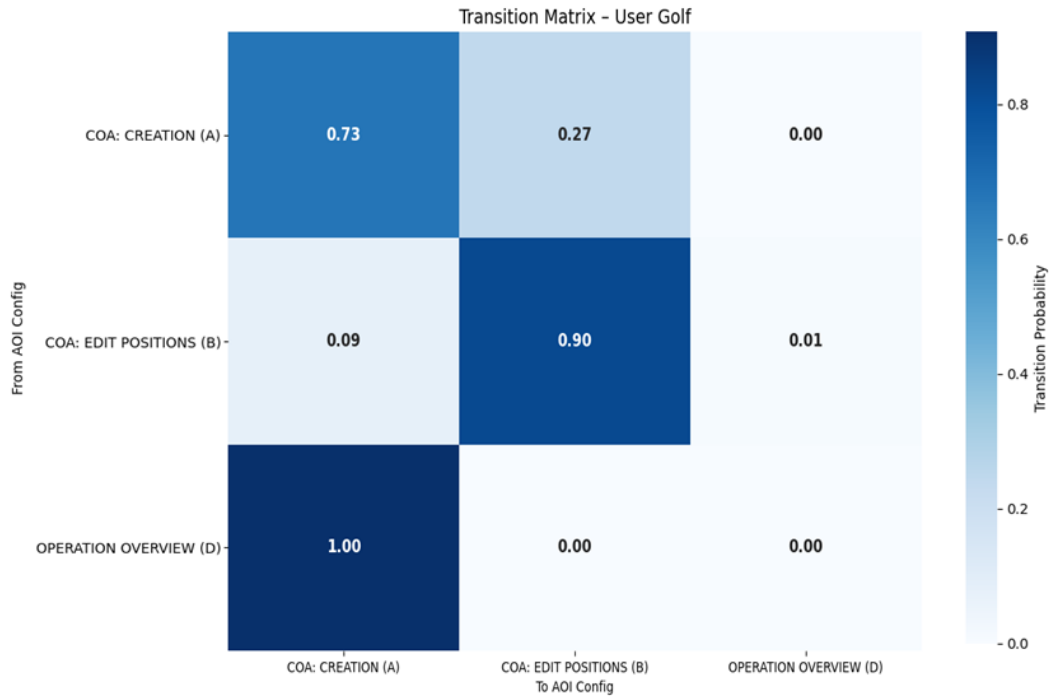


Figure 37. Session 3 transition matrix for Golf.

In planning session 3, user interaction patterns showed narrowing in focus with some nuanced differences. Alpha maintained a fairly balanced approach, continuing to cycle between Creation, Edit Positions, and Analysis with a notable portion of transitions emerging from Overview, suggesting a sustained emphasis on iterative scenario assessment. In contrast, Echo exhibited a sharper decline in cross-AOI movement with transitions becoming highly concentrated between Creation and Edit Positions and minimal activity in other modules. This reflects a tightening of engagement, possibly focusing on final execution. Golf preserved a narrow interaction profile but demonstrated a clear pattern of switching between Creation and Edit Positions, with Overview briefly visited and redirected to Creation. This reinforces a late-stage, execution-oriented strategy with limited diversification. Collectively, session 3 patterns suggest consolidation of planning behavior in some users and sustained adaptability in others, highlighting consistent individual differences in strategies across the full simulation period.

3.6 Cursor Movement Metrics

We analyzed pointer data to better understand user behavior and evolving cognitive states during the simulation, specifically movement straightness and cursor heatmaps across the three planning sessions. These metrics serve as valuable indicators of motor efficiency, task engagement, and potential cognitive depletion under varying task demands. Although gaze data typically provides more direct

insight into visual attention and cognitive processing, substantial missing and incomplete recordings in our dataset limited its utility. As a result, we focused on pointer-based measures, which prior research has shown to meaningfully reflect cognitive workload and interaction efficiency (Aoidh et al. 2012; Demšar and Çöltekin 2017). Furthermore, mouse-tracking data has been shown to approximate aspects of eye movements and attentional patterns (Freeman and Ambady 2010; Tahir et al. 2011, 2013), supporting its use as a practical proxy for visual engagement and cognitive state. Patterns such as curvilinear trajectories, spatial dispersion, and reduced movement straightness have been associated with increased mental workload and cognitive fatigue (Grimes and Valacich 2015), reinforcing the interpretive value of these measures.

Each figure in the following section captures pointer-based behavior across participants during simulation sessions. Visualizations include cursor heatmaps and pointer movement straightness metrics, offering insight into motor efficiency, task focus, and cognitive engagement over time. These indicators serve as practical proxies for attentional style and workload. At a high level, the data suggest distinct patterns:

- Alpha demonstrated the broadest and most variable pointer behavior, with widespread cursor dispersion and fluctuating straightness across AOIs, indicative of high interaction effort and iterative planning strategies.
- Echo showed moderate pointer efficiency with frequent shifts across AOIs, reflecting a flexible and adaptive approach. Heatmaps revealed exploratory activity, particularly in the Analysis interface.
- Golf consistently exhibited high pointer straightness and tightly clustered cursor paths, especially in Creation and Edit Positions, suggesting a structured and efficient interaction style with minimal exploration.

3.6.1 Session 1

Cursor heatmaps revealed substantial exploratory behavior across all users, particularly concentrated in the Edit Positions (B) AOI. This AOI consistently exhibited the most spatially dispersed and densely clustered activity, indicating that users engaged in more complex or frequent interactions within this workspace. In contrast, the Creation (A), Analysis (C), and Overview (D) AOIs displayed tighter cursor clusters, suggestive of more targeted, goal-directed behavior.

Despite these shared patterns, notable individual differences were observed (Figure 38):

- Alpha exhibited the broadest spatial cursor coverage in AOI B, with dense activity extending across nearly the entire central and lower interface

regions, indicating frequent and varied interactions. In contrast, cursor activity in AOIs A and C was more concentrated in the upper portion of the interface, suggesting more focused engagement. AOI D showed minimal interaction, with a small cluster of activity confined to the upper-left corner.

- Echo showed widespread cursor activity in AOI B, with dense interaction across the central interface. Compared with Alpha, this user also displayed stronger engagement in AOI C, particularly in the upper-left area. Activity in AOI A remained concentrated along the top, while AOI D showed a broader distribution, suggesting shifts in attention across interface areas.
- Golf showed a narrower interaction range overall. Their cursor activity in AOI B was dense but centrally focused, while AOI A revealed two distinct hotspots in the upper region of the interface. AOI C was not used at all, and AOI D showed a single, sharply concentrated area in the upper left. This pattern suggests a more focused and possibly streamlined interaction style.

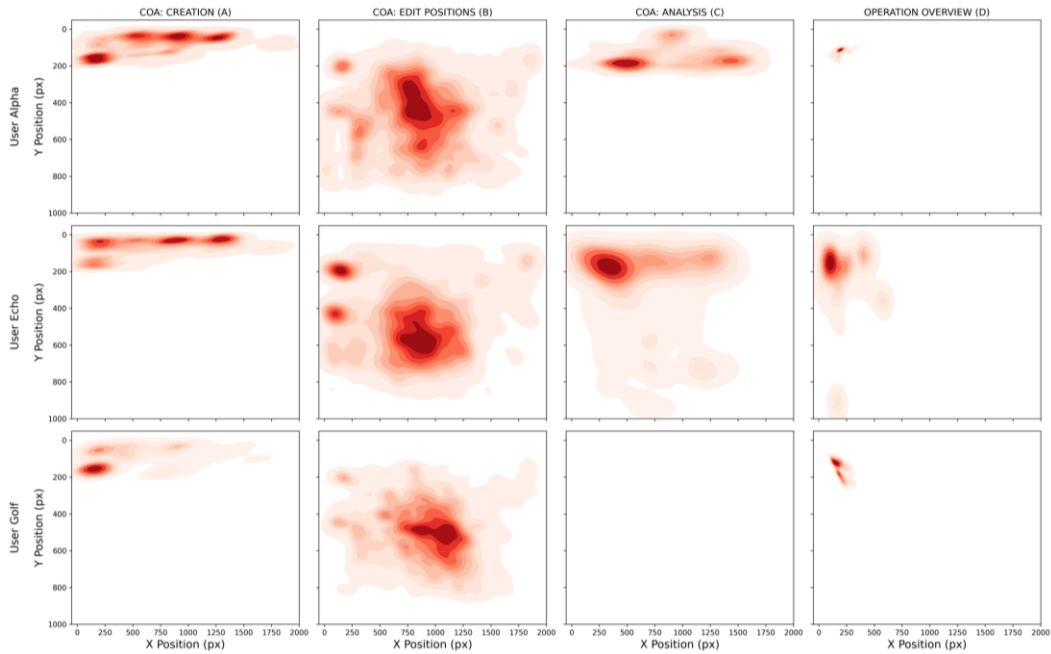


Figure 38. Session 1 cursor position heatmaps.

To quantify the efficiency of cursor navigation, we computed a pointer movement straightness metric for each user and AOI (Figure 39). Cursor data were segmented into nonoverlapping 10-s bins. Within each bin, straightness was calculated as the ratio of the direct (Euclidean) distance between the first and last cursor positions to the total path length traveled. Values approaching 1 indicate straight, goal-directed movement, while lower values reflect more circuitous or erratic patterns with frequent direction changes. Such movement inefficiencies have been associated

with increased cognitive workload, hesitation, or exploratory behavior (Grimes and Valacich 2015).

In session 1, all three users demonstrated variability in straightness across configurations:

- AOI C produced the highest straightness values for both Alpha and Echo, suggesting more efficient cursor movement, potentially due to task familiarity or reduced cognitive interference in that workspace.
- AOI B consistently yielded the lowest straightness across all users, with broader and flatter distributions indicative of exploratory or less efficient interaction patterns.
- Golf exhibited high straightness in AOI A, which may reflect a more structured or confident approach in that interface.
- AOI D resulted in lower straightness for Alpha and Echo but Golf showed moderately high values, again highlighting individual differences in interface engagement.

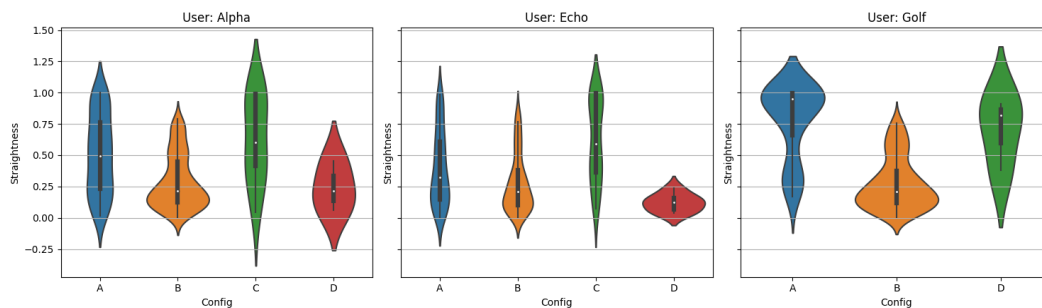


Figure 39. Session 1 cursor straightness distributions.

3.6.2 Session 2

In session 2, cursor heatmaps again revealed dominant interaction within the Edit Positions (B) AOI, especially for Alpha and Golf (Figure 40). The pointer activity in this AOI remained spatially distributed and dense, indicating high engagement and frequent task-switching. Compared with session 1, all users showed more involvement in the Analysis (C) AOI, suggesting either increased task complexity or deliberate re-evaluation of previously entered inputs. Creation (A) continued to exhibit interaction concentrated in the top interface region, consistent with prior UI element placements. Operation Overview (D) remained less frequently accessed but showed more extended use by Echo compared with others.

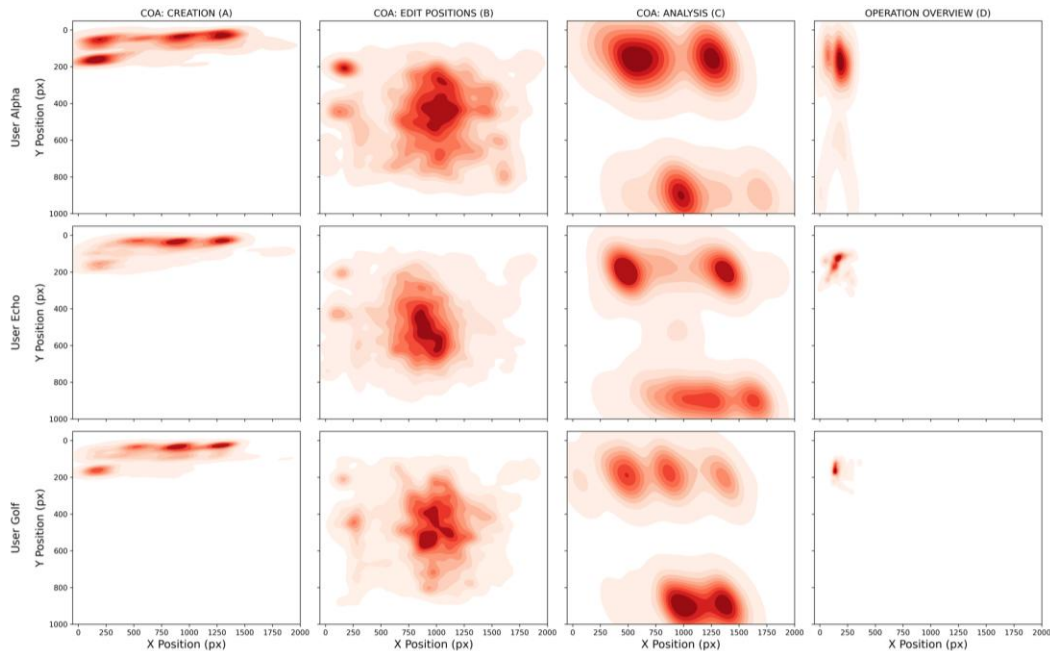


Figure 40. Session 2 cursor position heatmaps.

User-specific differences included the following:

- Alpha exhibited dense and sustained interaction within the Edit Positions (B) AOI, with activity extending broadly across both central and lower interface regions, suggesting continued reliance on this workspace for core interactions. In AOI C (Analysis), cursor activity became more structured compared with session 1, with clearly defined clusters in both upper and lower screen zones, potentially indicating focused attention on separate analytical panels. The Creation (A) AOI maintained top-bar interaction similar to session 1, while AOI D (Overview) again showed a narrow, upper-left engagement pattern, likely used for brief reference rather than ongoing task manipulation.
- Echo showed sustained, high-density interaction in AOI B (Edit Positions), centered tightly around the middle of the screen, reflecting consistent task engagement. AOI C (Analysis) revealed expanded engagement with multiple distinct hotspots, particularly across upper and lower regions, suggesting task-switching between summary views and detailed panels. Cursor activity in AOI A (Creation) remained concentrated along the top bar, similar to the previous day. Meanwhile, AOI D (Overview) demonstrated clustered activity in the upper-left.
- Golf maintained consistent patterns in AOIs A (Creation) and B (Edit Positions), with cursor activity again concentrated in the upper bar (A) and central interface (B), respectively. However, a notable change was observed

in AOI C (Analysis), previously unused on session 1, which now showed strong engagement with three distinct interaction zones, including both top-level summaries and detailed lower regions. This shift may reflect task progression or new analytical demands. AOI D (Overview) remained narrowly focused in the upper-left, suggesting quick, periodic referencing behavior.

To evaluate navigation efficiency on session 2, we again computed the pointer movement straightness metric for each user and configuration. As noted by Grimes and Valacich (2015), reduced straightness may be associated with increased cognitive load or exploratory behavior.

In session 2, straightness distributions revealed consistent patterns and individual differences across users (Figure 41):

- AOI C again yielded the highest straightness values for all users, especially Alpha and Echo, indicating that this AOI supported more efficient interactions.
- AOI B remained the least efficient across all participants, with broad, flat straightness distributions suggestive of sustained exploratory behaviors.
- Golf showed particularly high straightness in both AOIs A and C, reinforcing their focused interaction style.
- Echo displayed lower straightness in AOIs B and D, with narrow distributions centered near 0.2–0.3, again indicating less directed movement or greater task uncertainty in those interfaces.
- Alpha showed balanced straightness between AOIs A and C, while maintaining moderate variability in B and D, indicating some shifts in attention and engagement strategy.

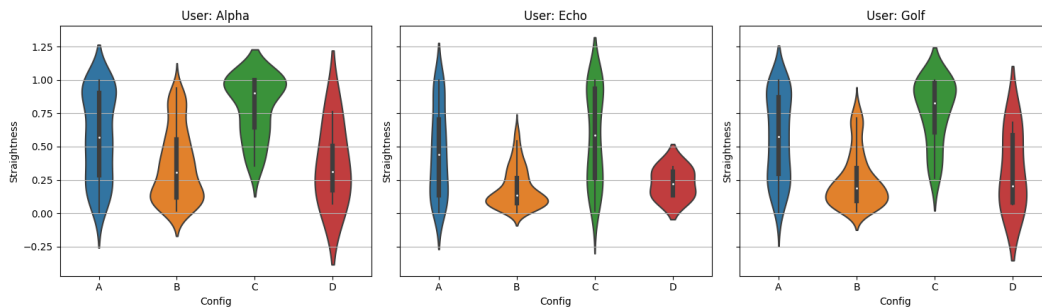


Figure 41. Session 2 cursor straightness distributions.

3.6.3 Session 3

In session 3, cursor heatmaps reflected a continuation of high engagement in the Edit Positions (B) AOI across all users, though with slightly reduced dispersion

compared with earlier days (Figure 42). This suggests that while AOI B remained the primary workspace. The Creation (A) AOI continued to show consistent activity along the top interface bar, suggesting habitual navigation patterns. Operation Overview (D) remained the least used but displayed compact, localized hotspots, indicating brief yet routine reference checks. Interaction within the Analysis (C) AOI appeared less prominent compared with session 2 with reduced intensity and narrower spatial coverage. This may reflect a reduced need for in-depth verification or a shift in analytic focus.

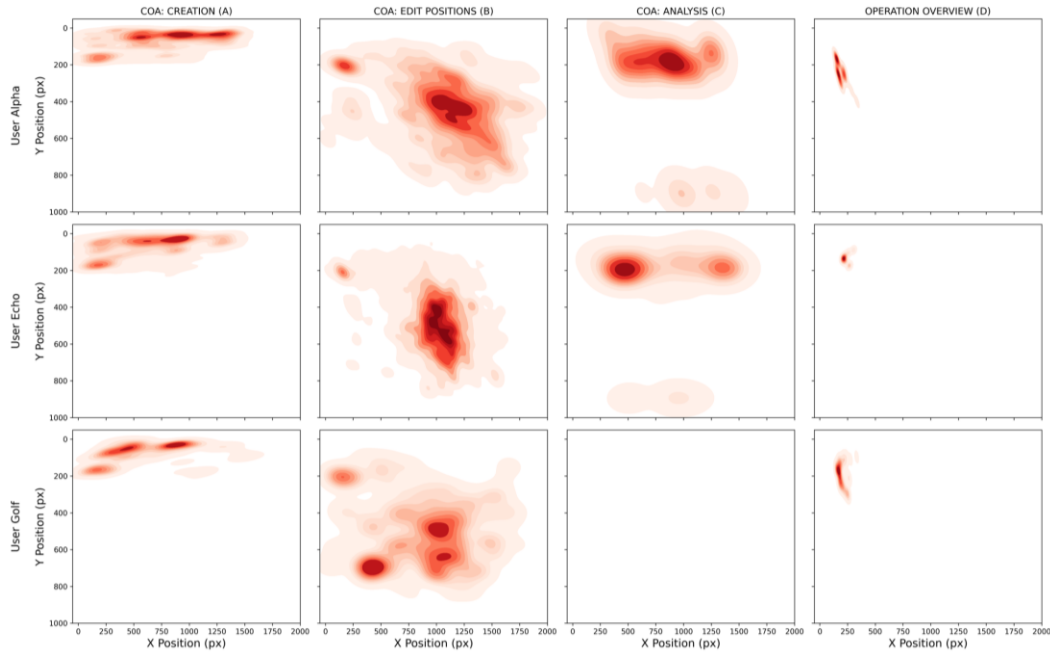


Figure 42. Session 3 cursor position heatmaps.

- Alpha maintained high-density interaction in AOI B, though with slightly more central clustering than in earlier days. AOI C showed a clear divergence, with upper and lower screen hotspots, suggesting toggling between high-level summaries and detailed elements. AOI A remained focused at the upper interface bar, while AOI D showed a sharper, diagonally structured cluster in the upper-left.
- Echo interaction patterns were consistent across AOIs, with AOI B showing dense activity in the center and AOI C displaying two primary hotspots in the upper interface. Compared with earlier days, AOI D use became more concentrated and less dispersed. Activity in AOI A remained tightly grouped along the top with no significant expansion.
- Golf continued to show dense engagement in AOI B, with concentrated cursor activity across the central workspace and parts of the lower interface. This pattern suggests consistent task execution and interaction in the Edit

Positions environment. AOI A remained active along the top UI bar, reflecting habitual navigation behavior. AOI C was not used at all, similar to session 1, indicating a sustained lack of interaction with the analysis interface. AOI D exhibited a tightly focused cluster in the upper-left corner.

To assess the efficiency of cursor navigation in session 3, we analyzed the pointer movement straightness metric across users and interface configurations (Figure 43).

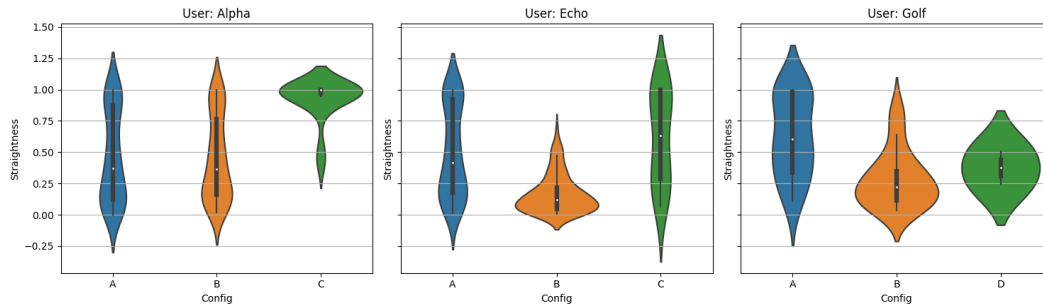


Figure 43. Session 3 cursor straightness distributions.

Across the three users, distinct straightness patterns emerged:

- Alpha displayed highly efficient and tightly clustered pointer movements in AOI C (Analysis) with values near 1. This suggests highly directed and confident cursor use. In contrast, AOIs A (Creation) and B (Edit Positions) showed lower and more variable straightness values, particularly for AOI B, indicating less efficient or more hesitant movements.
- Echo demonstrated high straightness in AOI C with a distribution skewed toward 1, suggesting strong directional control. AOI A showed moderate straightness with a wider spread, while AOI B had the lowest values with a flat and narrow distribution near the bottom of the scale, reflecting exploratory or effortful interaction.
- Golf showed consistent high straightness in AOI A (Creation), indicating structured, purposeful movement patterns. AOI D (Overview) exhibited moderately high straightness with less variability than observed on previous days. However, AOI B remained the least direct, showing a wide and flat distribution with many low straightness values.

Overall, AOI C once again produced the most efficient pointer movements for Alpha and Echo, mirroring the pattern seen in session 1 and suggesting ease of use with the analysis interface. AOI B remained the most interaction-heavy and least direct area across users, pointing to higher navigational demands or less structured tasks in that context. Golf's strong performance in AOI A reinforces their consistent and directed approach within the creation interface.

3.7 Eye-Tracking Metrics

To assess visual attention and cognitive dynamics during simulation tasks, eye-tracking data were collected using Tobii Pro Fusion systems (Tobii AB, Danderyd, Sweden), operating at a sampling frequency of 300 Hz. Eye-tracking contributes behavioral markers such as pupil size, fixation duration, blink frequency, and saccade dynamics, all of which have been shown to vary with cognitive demand (Skaramagkas et al. 2021; Mark et al. 2024; Ekin et al. 2025). The recorded signals included time-stamped gaze coordinates (horizontal and vertical positions) and pupil diameters for both eyes. These measurements were preprocessed and analyzed using a custom Python-based pipeline that segmented the data into overlapping 10-s windows updated every 1 s to align with mission events and preserve temporal resolution.

For each window, a comprehensive set of features was computed across four categories: pupil dynamics, blink activity, saccades, and fixations. The feature extraction approach followed established principles from the PyGaze toolkit (Dalmaijer et al. 2014), which provides validated methods for detecting oculomotor events and computing gaze-based behavioral metrics in dynamic task environments. A summary of the extracted eye-tracking features and their descriptions is provided in Table 6.

Table 6. Eye-tracking features extracted.

Feature category	Feature name	Description
Gaze saccades	Saccade count	Number of detected saccades per window
	Saccade velocity	Speed of saccadic eye movements
Gaze fixations	Fixation count	Number of fixations per window
	Fixation duration	Duration of gaze stabilizations on a single point
Pupillometry	Pupil size	Mean pupil diameter across both eyes
Blink activity	Blink count	Number of blinks, inferred from data gaps ≥ 300 ms

Due to a substantial amount of missing or incomplete eye-tracking data in sessions 2 and 3, given the participants' movement between their stations during collaborative planning, our analysis focused exclusively on session 1, where data quality and temporal coverage were sufficient to support reliable feature extraction and interpretation.

Figure 44 presents findings for these features, offering insight into users' visual scanning strategies and cognitive engagement. Although data analyses were limited to session 1, clear differences in eye-movement profiles emerged across participants.

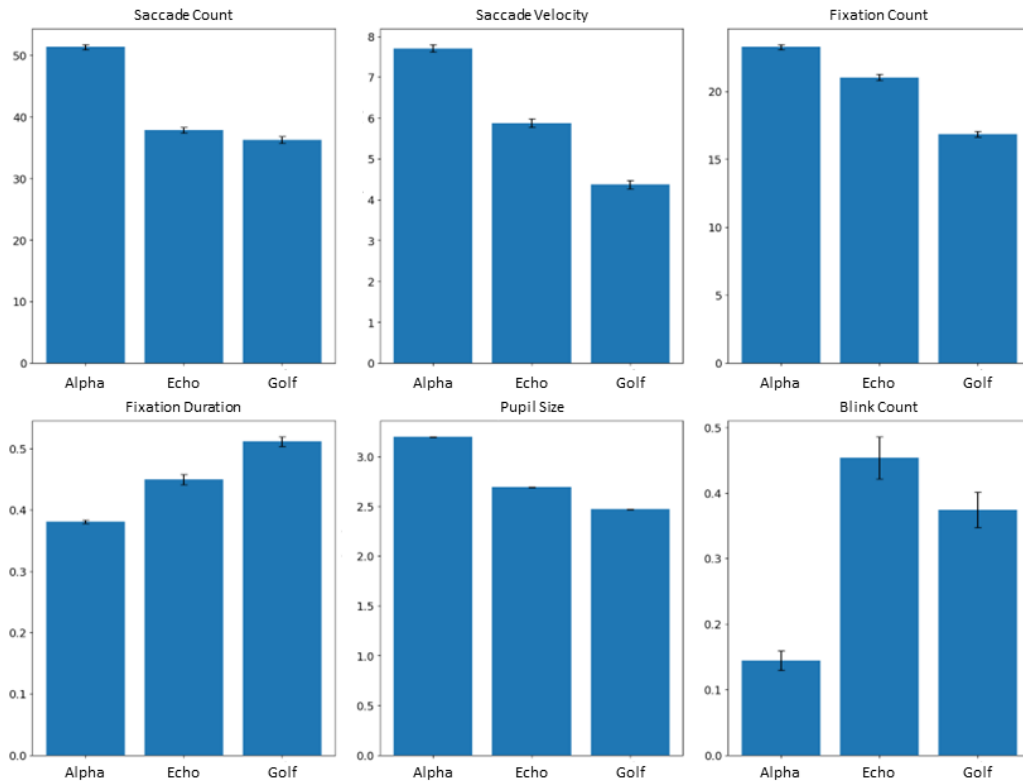


Figure 44. Select eye features.

Alpha demonstrated the most visually active profile. High saccade count and saccadic velocity variability indicate frequent, rapid shifts in gaze, suggesting a broad visual scanning strategy. This is further supported by the highest fixation count and shortest fixation durations, reflecting fast-paced attention shifts across interface elements. Alpha also showed the largest pupil size, a physiological marker commonly associated with elevated cognitive workload (Marquart et al. 2015). The low blink rate reinforces this interpretation, suggesting continuous attentional engagement. Altogether, the eye-tracking profile of Alpha reflects high visual vigilance and active monitoring behavior. Behaviorally, Alpha spent approximately 86% of the session in Edit Positions (AOI B) and 13% in Creation (AOI A), with regular transitions between the two. Cursor heatmaps revealed broad spatial coverage in AOI B, and pointer movement within that AOI showed the lowest straightness, reflecting exploratory and adaptive interaction. Taken together, the physiological and behavioral patterns suggest a focused but high-effort engagement style.

Echo demonstrated a moderately active visual profile. Saccade count and saccadic velocity variability were within the midrange across participants, indicating a balanced pattern of gaze shifts; neither overly frequent nor overly narrow. Fixation count was moderate, with fixation durations slightly longer than those of Alpha,

suggesting more sustained attention on individual interface elements. Pupil size was smaller than Alpha's, and blink rate was the highest among the group; both markers typically associated with reduced or fluctuating cognitive workload. Together, these patterns suggest a measured but less sustained engagement strategy, with periodic attentional breaks that may reflect a more flexible or reactive visual monitoring approach. Unlike Alpha's narrower focus, Echo demonstrated broader interface engagement, spending 68% of the session in Edit Positions, with substantial time in Creation, and additional use of Analysis and Overview. The participant made frequent transitions across AOIs, and cursor data showed spatially distributed interactions and moderate straightness, particularly in AOI C. Together, these patterns indicate a strategically flexible planning style, characterized by broad situational exploration and more intermittent, paced cognitive engagement; less visually intense than Alpha, but more diversified in interaction.

Golf showed the least active profile of the three. Saccade count and saccadic velocity variability were both low, indicating minimal gaze shifts and a narrow visual field. Fixation count was relatively low, and fixation durations were the longest among the group, suggesting a slower attentional rhythm with sustained focus on fewer interface elements. Pupil diameter was smaller than average, and blink rate was slightly elevated; both markers typically associated with lower cognitive workload and reduced engagement. The eye-tracking profile of Golf suggests a more passive and constrained visual approach, potentially reflecting reduced task demand or a more simplified interaction strategy. These visual patterns are further corroborated by behavioral data: the participant spent approximately 97% of the session in the Edit Positions interface (AOI B), with virtually no transitions to other AOIs, indicating a limited task flexibility. Pointer movement exhibited relatively high straightness across AOIs, particularly in AOI A, reflecting linear navigation with minimal course correction. Taken together, the lack of visual variability, limited interface engagement, and highly directed motor behavior point to a low cognitive workload profile, characterized by focused but inflexible task execution and minimal adaptive exploration.

4. Conclusion

This touchpoint successfully demonstrated the value of a prototype software system, developed and integrated by the HCxS 6.2 HE research team in support of SC3 program for rapid, streamlined COA development, prediction, and refinement. The assessment of the MCBL Soldier participants and our analysis of their use of the ITS indicate it will dramatically increase the pace of planning when fully matured and deployed in a future, technology-enabled mobile C2 nodes. In addition to demonstrating the clear value and future impact of this system, the touchpoint

provided the research team with a wealth of information to guide future improvements in the technology itself, team experimentation methodologies, and nonmateriel approaches to C2 for the Army in the FOE.

5. References

- Aoidh EM, Bertolotto M, Wilson DC. Towards dynamic behavior-based profiling for reducing spatial information overload in map browsing activity. *Geoinformatica*. 2012;16:409–434.
- Carleton RN, Norton MPJ, Asmundson GJ. Fearing the unknown: a short version of the Intolerance of Uncertainty scale. *Journal of Anxiety Disorders*. 2007;21(1):105–117.
- Dalmaijer ES, Mathôt S, Van der Stigchel S. PyGaze: an open-source, cross-platform toolbox for minimal-effort programming of eyetracking experiments. *Behavior Research Methods*. 2014;46:913–921.
- Demšar U, Çöltekin A. Quantifying gaze and mouse interactions on spatial visual interfaces with a new movement analytics methodology. *PloS One*. 2017;12(8):e0181818.
- Ekin M, Krejtz K, Duarte C, Duchowski AT, Krejtz I. Prediction of intrinsic and extraneous cognitive load with oculometric and biometric indicators. *Scientific Reports*. 2025;15(1):5213.
- Freeman JB, Ambady N. MouseTracker: software for studying real-time mental processing using a computer mouse-tracking method. *Behavior Research Methods*. 2010;42(1):226–241.
- Goecks VG, Waytowich N. COA-GPT: generative pre-trained transformers for accelerated course of action development in military operations. In: 2024 International Conference on Military Communication and Information Systems (ICMCIS); 2024 Apr 23–24; Koblenz, Germany. IEEE; 2024. p. 1–10. <https://doi.org/10.1109/ICMCIS61231.2024.10540749>
- Grimes M, Valacich J. Mind over mouse: the effect of cognitive load on mouse movement behavior. *Proceedings of the 36th International Conference on Information Systems (ICIS) - Exploring the Information Frontier*; 2015 Dec 13–16; Fort Worth, TX.
- Hamilton K, Shih SI, Mohammed S. The development and validation of the rational and intuitive decision styles scale. *Journal of Personality Assessment*. 2016;98(5):523–535. <https://doi.org/10.1080/00223891.2015.1132426>
- Headquarters, Department of the Army. Planning and orders production. Headquarters, Department of the Army; 2024. Field Manual No.: FM 5-0.

- https://armypubs.army.mil/epubs/DR_pubs/DR_a/ARN44590-FM_5-0-001-WEB-3.pdf
- Mark JA, Curtin A, Kraft AE, Ziegler MD, Ayaz H. Mental workload assessment by monitoring brain, heart, and eye with six biomedical modalities during six cognitive tasks. *Frontiers in Neuroergonomics*. 2024;5:1345507.
- Marquart G, Cabrall C, De Winter J. Review of eye-related measures of drivers' mental workload. *Procedia Manufacturing*. 2015;3:2854–2861.
- Mathieu JE, Luciano MM, D'Innocenzo L, Klock EA, LePine JA. The development and construct validity of a team processes survey measure. *Organizational Research Methods*. 2020;23(3):399–431.
- Schneider TR et al. The influence of trust propensity on behavioral trust. Paper presented at: Association for Psychological Society; 2017 May 25–28; Boston, MA.
- Skaramagkas V et al. Review of eye tracking metrics involved in emotional and cognitive processes. *IEEE Reviews in Biomedical Engineering*. 2021;16:260–277.
- Tahir A, McArdle G, Bertolotto M. A web-based visualisation tool for analysing mouse movements to support map personalisation. In: *International Conference on Database Systems for Advanced Applications*. Springer Berlin Heidelberg; 2011 April. p. 132–143.
- Tahir A, McArdle G, Bertolotto M. Comparing close destination and route-based similarity metrics for the analysis of map user trajectories. In: *Web and Wireless Geographical Information Systems: 12th International Symposium, W2GIS 2013*; 2013 Apr 4–5; Banff, Canada. Springer Berlin Heidelberg; c2013. p. 117–128.

Appendix. Supporting Material*

* This appendix appears in its original form, without editorial change.

A.1. Supporting Materials

The following supporting materials represent the subsets of the collected/inspected data that were employed to inform the synthesized findings reported in the main body of this report. Each supporting material section includes the subset of data mapped to each of the learning demands, sub-question, or set of sub-questions such that the reader may more readily understand the set of information employed to support each separately. The full set of surveys administered and data captured in service of the May 2025 SC3 touchpoint may be found in other sections of this appendix. Note that, for navigating the supporting material below, each subset of mapped materials are separated by a horizontal line, followed by the learning demand, sub-question, or set of sub-questions addressed by the supporting materials and then the supporting materials themselves

A.1.1 Usability and Features

A.1.1.1 High-level ITS Feedback Cluster

SC3 Touchpoint Session 3 ITS Evaluation (item 7)

Finally, please provide any other feedback you would like to see regarding the ITS, the touchpoint sessions, the general direction of the research effort, or anything we have not touched on this week. [session 3]

Alpha:

“I think the research is going in the right direction. For future touchpoints we need to make sure we have all the components necessary for a wargame (Friendly COA, Enemy COA, maybe add COA Evaluation criteria). This will allow us to better see how the simulation could simulate a wargame somewhat similar to the traditional method. There is still a lot of areas for continued discussion on how to best simulate a wargame with outputs that enable staff and commander decisions. The system has a lot of potential, and I look forward to continuing working with you all to refine the simulation and COA GPT.”

Echo:

“I would not use this tool right now to execute MDMP in an exercise or a live event, but neither is the maturity of the tool to be expected to do so. There is a great deal of promise in the ability to communicate between COA-GPT and a simulator to then run through the tasks that COA-GPT has defined. The further refinement of how to make those changes in COA-GPT

and its ability to then push them to the simulator in a way that is usable is essential to provide a more cohesive, comprehensive product.”

Golf:

“We touched on everything during our discussions. Once [capabilities for sketch and natural language interaction] gets on board I think the ability to see the graphics will happen. That will be a big step.

I want to eventually see COA-GPT look at the higher’s mission and then recommend different task organizations (create different task forces) in order to execute the mission.”

SC3 Touchpoint Session 3 ITS Evaluation (item 4)

What impact do you think the full, combined ITSs outputs, results, outcomes, metrics, etc. had on your planning process compared to traditional COA development? [session 3]

Alpha:

“I was able to get to the different phases of MDMP faster; however, right now the outputs do not provide the clarity required to have confidence in the outputs meeting the same standard as the traditional COA development carried out by a trained staff.”

Echo:

“The full combined efforts are significant in the fact that COA-GPT has taken a series of information inputs through Mission analysis and adapted that information into a COA Sketch. The linkages between the simulation and COA-GPT enabled testing of those defined COAs. The ability to manipulate the COA using COA-GPT was limited in that changes did not occur like I wanted and there were limitations in potentially the proper way to prompt COA-GPT to get it to function the way I wanted but at the same time COA-GPT told me it had accomplished tasks that upon further investigation is showed inaccuracies in the sketch it provided.”

Golf:

“SC3 and COA-GPT are doing great work for where we began from less than a year ago. I’m confident that the SC3 sim combined with COA-GPT will continue to get better. It’s not perfect now, but even now, it’s a good start point.”

SC3 Touchpoint Session 3 ITS Evaluation (item 5)

What interface or visualization feature do you think contributed most to your feeling of whether you could trust and make use of the seed COAs as visualized in the simulator? (note: that could be a positive or negative impact)

Alpha:

“Currently a few units made their way across the Brigade boundary into another Brigade’s Area of operation. It also does not have a clear way currently to display how each unit accomplished the different phases of the operation.

One positive, the adjustments made in COA GPT for 1-1CAV were somewhat reflected in the SIM by updating the Phase lines it would defend. It put them in the proper sequencing.”

Echo:

“I think the initial seed COAs combined with the running of the simulation with the tasks assigned has a lot of promise. The ability to manipulate the simulation was easy to use and once on the simulator it allowed for updated COAs much more streamlined. The reconfiguring of the COA-GPT product took some time as I can only assume that it is trying to synchronize across many products but the COA-GPT sketch still struggles to adapt to changes. Again this might be how to describe the change to COA-GPT for it to understand as I even attempted to provide grid references for some things which did not produce any results and I am not sure why that didn’t work as the prompt said that it did.”

Golf:

“The most negative feature of the simulator was the glob of rings the system defaults to (glob of blue and red circles). I think it should default to just the control measure graphics and unit icons.

I like how COA-GPT allowed me to click on and off the phase graphics.”

A.1.1.2 Useful Features

Questionnaire for User Interface Satisfaction (item 7)

Please list the most positive aspects of the ITS (identify 3) [session 1]

Alpha:

“Easy to move the units on the screen.

Easy to add a new task to the list.

Ability to see range of weapon systems on the map”

Echo:

“Easily depicts the routes after the simulation has run.

Identifies deaths of friendlies and enemies automatically as well as some status bars.

Ease of selection of COA and the point and click functionality to understand prescribed tasks”

Golf:

“#1 Easy to use.

#2 I liked how I could continually run the simulation (after making just a few inputs/changes).”

#3 The model was easy to learn and understand”

System Elements Evaluation (item 1)

What Integrated Technology System (ITS) features were useful and how? [session 2] Please try to comment on the specific elements listed below: What specifically about the COA-GPT component? What specifically about the constructive simulation component? What specifically about the user interface component?

Alpha:

“COA GPTs ability to take large volumes of data and quickly pull-out key pieces of information is very valuable.

The constructive simulation’s ability to quickly process a COA and provide analysis has the potential to speed up the wargaming process.

The user interface concept to create shared understanding across dispersed command nodes through an augmented reality map has great potential to improve the operation processes.”

Echo:

“The ability to seed the simulation with the baseline information was key to avoid having to drag and drop elements and then build out as needed.

The baseline of having a point and click interface to direct moving units across a map with known terrain features was helpful in providing some reference points for action.”

Golf:

“COA GPT is a tool that continues to get better and better. The simulation’s ability to move units and give the unit a task is already impressive.”

A.1.1.3 Problematic Features

Questionnaire for User Interface Satisfaction (item 6)

Please list the most negative aspects of the ITS (identify 3) [session 1]

Alpha:

“Adding tasks didn’t allow you to insert between two existing tasks.

Was not clear which location icon related with which task.

When adding a subsequent task the location icon went back to unit rather than previous task”

Echo:

“Hard to understand decisions based of unit types/difference in unit types and their capabilities. Inability to compare the COAs (projected) but manual process of comparing COAs is not viable every time you change the analysis is not saved. Validation that each element has a task is difficult to discern from the interface”

Golf:

“#1 Couldn’t turn off the route of units... I don’t want to see all unit routes. If I’m planning this operation, I don’t care where the 303 Psyop company is going... or the Medical company.

#2 When I go to a different editor and then return to where I was... all of the boxes are checked... so I have to go back un check Unit ranges or terrain.

#3 Sometimes I want to move an entire battalion icon at once rather than every (4x) company”

System Elements Evaluation (item 2)

What ITS features were problematic and how? [session 2] Please try to comment on the elements listed below: What specifically about the COA-GPT component? What specifically about the constructive simulation component? What specifically about the user interface component

Alpha:

“Right now, COA GPTs ability to take terrain into consideration significantly limits its ability to provide thorough COA generation.

constructive simulation: Not being able to see the operations broken down into phases and coordinate the operations in time and space is limiting. the inability to accurately simulate indirect fire prevents a complete understanding of the COA’s effectiveness”

Echo:

“The user interface does need some work as the ability to work between multiple COAs should be saved and analysis something that could be referenced without having to take external notes. Being able to compare between one COA to another requires saving that data for further use. The ability have distinct COAs provided that directly tie to the order is also necessary as COA 4 in its description did not allow for fixing of Monroe which was a directed task. Interface wise the need to be able to know what each unit is doing will become more complex the more units are on the page. A way to aggregate headquarters with their action arms to consolidate tasks and move together would be helpful instead of individual entities at the company level. This would help in providing the functionality of having a headquarters and the supporting elements as part of the formation”

Golf:

“I assume (could be wrong) COA-GPT wrote the COAs. If yes, COA-GPT did not provide viable COAs (ones that could be presented to a COL level commanders. The COA i looked at didn’t mention Objective MONROE and MONROE is the division’s main effort.

Simulation & its interface: It’s new and for being new it does a great job. We’d like to be able to de-clutter range rings. Units in the assembly area need to be co-located... i.e. all 1-35 Armor units should be relatively close together”

A.1.1.4 Lacking Features

System Elements Evaluation (item 3)

Which element of the ITS is most lacking, specifically in terms of hindering your ability to conduct planning in a sufficiently realistic manner (e.g., gaps in representations, oversimplifications, inaccuracies, etc.)? [session 2] Please

consider what you would say are the most critical of these lacking areas to address and in what way would you like them addressed?

Alpha:

“Right now, COA GPTs ability to take terrain into consideration significantly limits its ability to provide thorough COA generation.

Not being able to see the operations broken down into phases and coordinate the operations in time and space is limiting.”

Echo:

“The most lacking of these elements is the functionality of the systems inside the simulation and how that then effects the COA analysis. The tasking of units only seems to be focused on named elements inside the COA narrative but no sequencing seems to be in thought as the simulation is drafted. Sequencing of all of the elements of a combined arms engagement is key to how the Army functions to allow for the full power. Identifying that main effort and supporting efforts while continuing to differentiate functions of those units is necessary to gain trust in the realism of the simulation to provide an adequate enough representation of the fight to use as a test of the actual engagement.”

Golf:

“There are many icons on the map. I’d like to be able to find a specific unit’s location (Ctrl F). I couldn’t find several enemy units that were templated.

If possible, I’d like to see the operation’s timeline on the map. This would help to sequence events... as stated earlier, the PAAs need to be cleared before the brigade seizes Obj MADISON. 4-27 Arty battalion needs to move into the cleared PAAs before the brigade seizes Obj MADISON.”

A.1.1.5 Planning Effects

SC3 Touchpoint Session 3 ITS Evaluation (item 4)

What impact do you think the full, combined ITSs outputs, results, outcomes, metrics, etc. had on your planning process compared to traditional COA development? Please consider commenting on the speed and perceived usefulness of SC3 Sim’s feedback compared to traditional methods of estimating COA outcomes.” [session 3]

Alpha:

“I was able to get to the different phases of MDMP faster; however, right now the outputs do not provide the clarity required to have confidence in the outputs meeting the same standard as the traditional COA development carried out by a trained staff.”

Echo:

“The full combined efforts are significant in the fact that COA-GPT has taken a series of information inputs through Mission analysis and adapted that information into a COA Sketch. The linkages between the simulation and COA-GPT enabled testing of those defined COAs. The ability to manipulate the COA using COA-GPT was limited in that changes did not occur like I wanted and there were limitations in potentially the proper way to prompt COA-GPT to get it to function the way I wanted but at the same time COA-GPT told me it had accomplished tasks that upon further investigation is showed inaccuracies in the sketch it provided.”

Golf:

“SC3 and COA-GPT are doing great work for where we began from less than a year ago. I’m confident that the SC3 sim combined with COA-GPT will continue to get better. It’s not perfect now, but even now, it’s a good start point”

SC3 Touchpoint Session 2 ITS Evaluation (item 8)

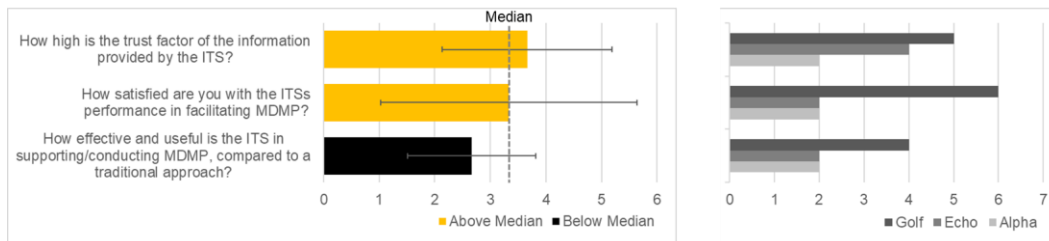


Figure A-1. Session 2 ITS evaluation.

Table A-1. Session 2 ITS evaluation.

Item query	Aggregate ($\mu \pm \sigma$)	Alpha	Echo	Golf
How high is the trust factor of the information provided by the ITS?	3.67 ± 1.53	2	4	5
How satisfied are you with the ITSs performance in facilitating MDMP?	3.33 ± 2.31	2	2	6
How effective and useful is the ITS in supporting/conducting MDMP, compared to a traditional approach?	2.67 ± 1.15	2	2	4

SC3 Touchpoint Session 1 ITS Evaluation (items 1-16)

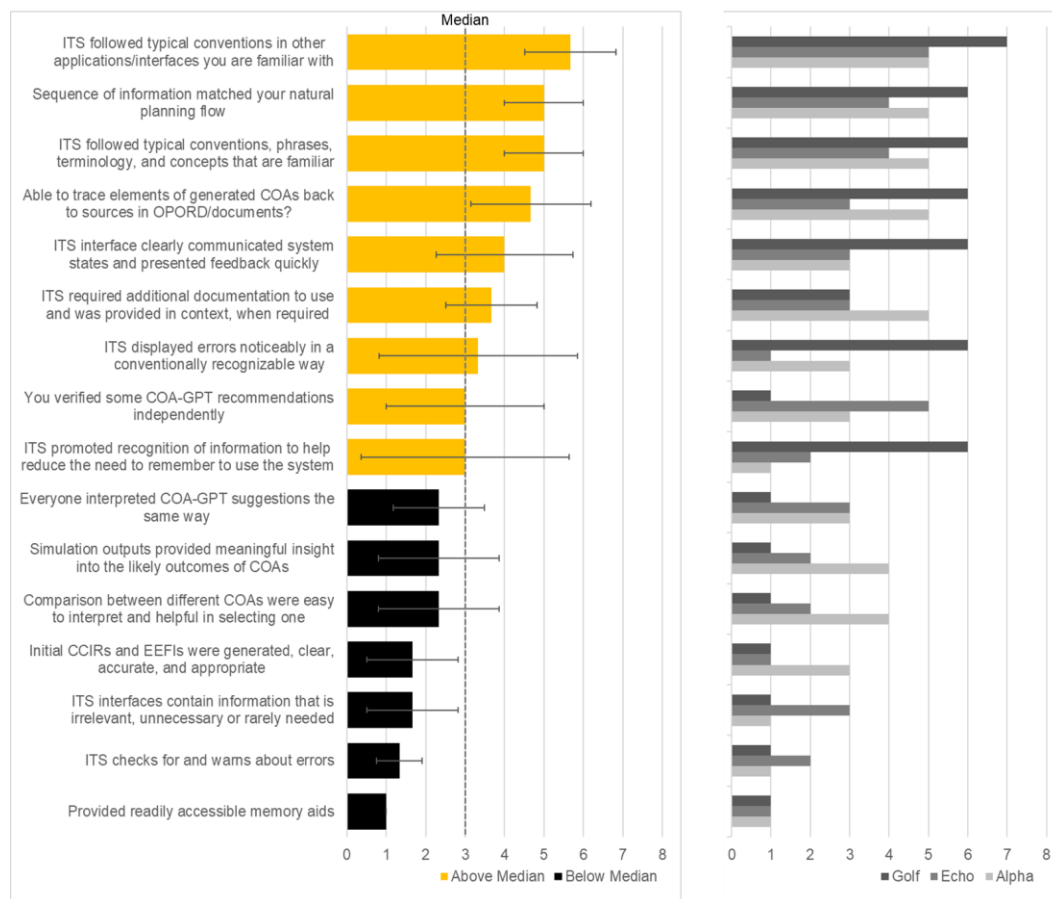


Figure A-2. Session 1 COA-GPT evaluation.

Table A-2. Session 1 COA-GPT evaluation.

Item query	Aggregate ($\mu \pm \sigma$)	Alpha	Echo	Golf
Did the ITS follow typical conventions seen in other applications or interfaces you are familiar with	5.67 ± 1.15	5	5	7
Did the ITS follow typical/standard conventions, phrases, terminology, and concepts that are familiar to you?	5.00 ± 1.00	5	4	6
Did the sequence of information match your natural planning flow?	5.00 ± 1.00	5	4	6
Were you able to trace elements of the generated COA back to their sources in the original operations order or input documents?	4.67 ± 1.53	5	3	6
Did the ITS interface clearly communicate system states and present feedback quickly	4.00 ± 1.73	3	3	6
Did the ITS require additional documentation or explanation to use and, if so, was the documentation provided in context, at the moment you required it?	3.67 ± 1.15	5	3	3
Did the ITS display errors messages noticeably in a conventionally recognizable way?	3.33 ± 2.52	3	1	6
Did the ITS promote recognition of information to help you reduce the need to remember to use the system (e.g., Strikethrough often uses the icon to easily recognize the function)?	3.00 ± 2.65	1	2	6
Did you verify any of COA-GPTs recommendations independently? What prompted that decision?	3.00 ± 2.00	3	5	1
Was the comparison between different COAs easy to interpret and helpful in selecting one?	2.33 ± 1.53	4	2	1
Did the simulation outputs provide meaningful insight into the likely outcomes of each COA?	2.33 ± 1.53	4	2	1
Did everyone interpret the COA-GPTs suggestions the same way?	2.33 ± 1.15	3	3	1
Did the ITS interfaces contain information that is irrelevant, unnecessary or rarely needed?	1.67 ± 1.15	1	3	1
Were Initial CCIRs and EEFI generated and, if so, were they clear, accurate, and appropriate for the mission-context?	1.67 ± 1.15	3	1	1
Did the ITS check for and warn you about errors (e.g., settings not saved or form fields left empty)?	1.33 ± 0.58	1	2	1
Did the ITS provide readily accessible memory aids?	1.00 ± 0.00	1	1	1

SC3 Touchpoint Session 2 ITS Evaluation (item 1 (a-d))

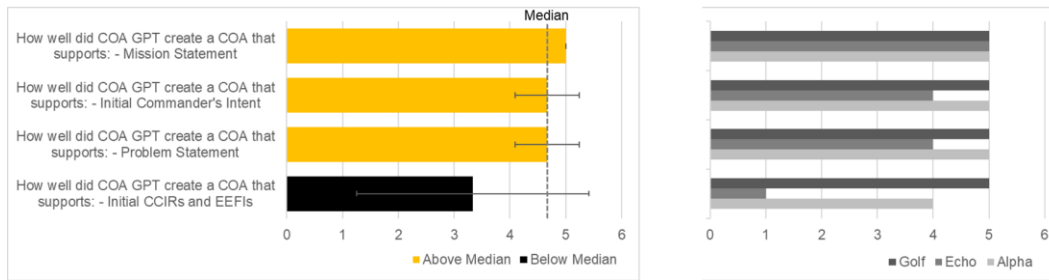


Figure A-3. Session 2 COA-GPT evaluation.

Table A-3. Session 2 COA-GPT evaluation.

Item query	Aggregate ($\mu \pm \sigma$)	Alpha	Echo	Golf
How well did COA GPT create a COA that supports: - Mission Statement	5.00 $\pm$ 0.00	5	5	5
How well did COA GPT create a COA that supports: - Initial Commander's Intent	4.67 $\pm$ 0.58	5	4	5
How well did COA GPT create a COA that supports: - Problem Statement	4.67 $\pm$ 0.58	5	4	5
How well did COA GPT create a COA that supports: - Initial CCIRs and EEfIs	3.33 $\pm$ 2.08	4	1	5

SC3 Touchpoint Session 2 ITS Evaluation (item 4 (a-k only))

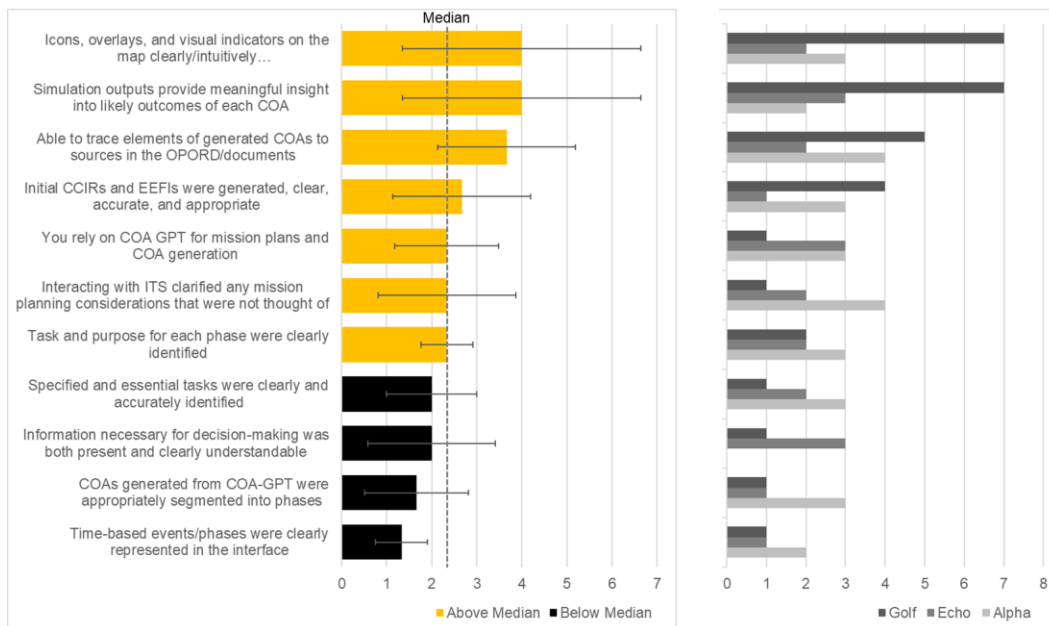


Figure A-4. Session 2 ITS evaluation.

Table A-4. Session 2 ITS evaluation.

Item query	Aggregate ($\mu \pm \sigma$)	Alpha	Echo	Golf
Did the simulation outputs provide meaningful insight into the likely outcomes of each COA?	4.00 ± 2.65	2	3	7
Were icons, overlays, and other visual indicators on the map (such as those associated with various units, terrain, Avenues of Approach (AoA), or Defensive positions) clearly and intuitively distinguishable, and doctrinally appropriate?	4.00 ± 2.65	3	2	7
Were you able to trace elements of the generated COA back to their sources in the original operations order or input documents?	3.67 ± 1.53	4	2	5
Were Initial CCIRs and EEFIIs generated and, if so, were they clear, accurate, and appropriate for the mission-context?	2.67 ± 1.53	3	1	4
Did interacting with ITS clarify any mission planning considerations you hadn't thought of?	2.33 ± 1.53	4	2	1
To what extent would you rely on COA GPT for mission plans and COA generation?	2.33 ± 1.15	3	3	1
Was the task and purpose for each phase clearly identified?	2.33 ± 0.58	3	2	2
Was the information necessary for decision-making both present and clearly understandable?	2.00 ± 1.41	0	3	1
Were Specified and Essential Tasks clearly and accurately identified?	2.00 ± 1.00	3	2	1
Was the COA generated from COA-GPT appropriately segmented into phases	1.67 ± 1.15	3	1	1
Were time-based events and phases clearly represented in the interface (e.g., operation start, phase shifts, timing of unit actions)?	1.33 ± 0.58	2	1	1

A.1.1.6 Trust and Reliance

Human-Computer Trust (short) (items 1-15)

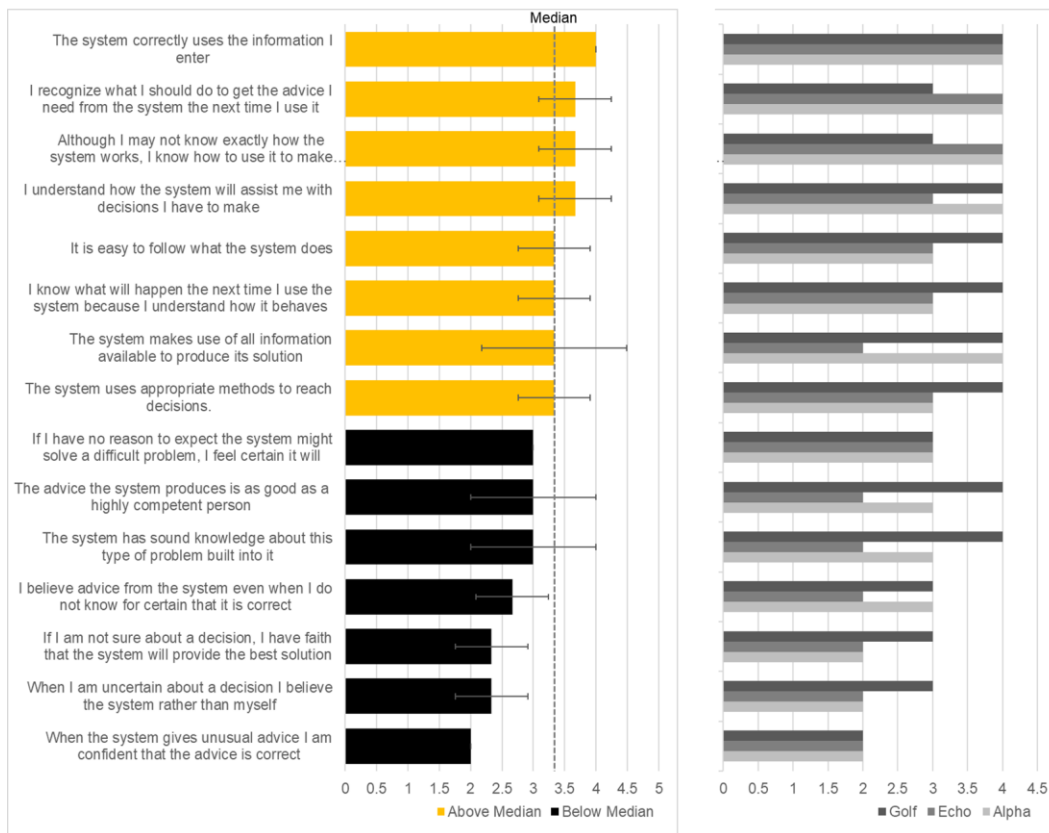


Figure A-5. Session 1 human-computer trust scale (short).

Table A-5. Session 1 human-computer trust scale (short).

Item query	Aggregate ($\mu \pm \sigma$)	Alpha	Echo	Golf
The system correctly uses the information I enter	4.00 ± 0.00	4	4	4
I understand how the system will assist me with decisions I have to make	3.67 ± 0.58	4	3	4
Although I may not know exactly how the system works, I know how to use it to make decisions about the problem	3.67 ± 0.58	4	4	3
I recognize what I should do to get the advice I need from the system the next time I use it	3.67 ± 0.58	4	4	3
The system makes use of all the knowledge and information available to it to produce its solution to the problem	3.33 ± 1.15	4	2	4
The system uses appropriate methods to reach decisions.	3.33 ± 0.58	3	3	4
I know what will happen the next time I use the system because I understand how it behaves	3.33 ± 0.58	3	3	4
It is easy to follow what the system does	3.33 ± 0.58	3	3	4
The system has sound knowledge about this type of problem built into it	3.00 ± 1.00	3	2	4
The advice the system produces is as good as that which a highly competent person could produce	3.00 ± 1.00	3	2	4
Even if I have no reason to expect the system will be able to solve a difficult problem, I still feel certain that it will	3.00 ± 0.00	3	3	3
I believe advice from the system even when I don't know for certain that it is correct	2.67 ± 0.58	3	2	3
When I am uncertain about a decision I believe the system rather than myself	2.33 ± 0.58	2	2	3
If I am not sure about a decision, I have faith that the system will provide the best solution	2.33 ± 0.58	2	2	3
When the system gives unusual advice I am confident that the advice is correct	2.00 ± 0.00	2	2	2

Human-Computer Trust (short) (items 1-15)

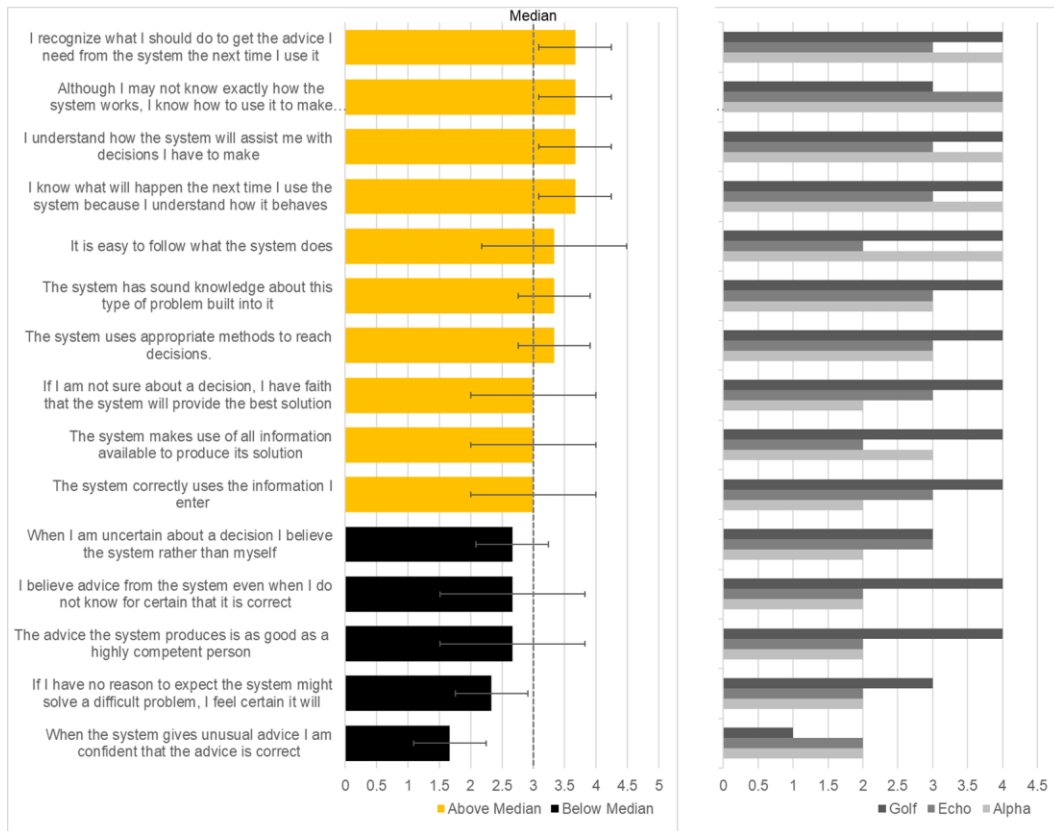


Figure A-6. Session 2 human-computer trust scale (short).

Table A-6. Session 2 human-computer trust scale (short).

Item query	Aggregate ($\mu \pm \sigma$)	Alpha	Echo	Golf
I know what will happen the next time I use the system because I understand how it behaves	3.67 ± 0.58	4	3	4
I understand how the system will assist me with decisions I have to make	3.67 ± 0.58	4	3	4
Although I may not know exactly how the system works, I know how to use it to make decisions about the problem	3.67 ± 0.58	4	4	3
I recognize what I should do to get the advice I need from the system the next time I use it	3.67 ± 0.58	4	3	4
It is easy to follow what the system does	3.33 ± 1.15	4	2	4
The system uses appropriate methods to reach decisions.	3.33 ± 0.58	3	3	4
The system has sound knowledge about this type of problem built into it	3.33 ± 0.58	3	3	4
The system correctly uses the information I enter	3.00 ± 1.00	2	3	4
The system makes use of all the knowledge and information available to it to produce its solution to the problem	3.00 ± 1.00	3	2	4
If I am not sure about a decision, I have faith that the system will provide the best solution	3.00 ± 1.00	2	3	4
The advice the system produces is as good as that which a highly competent person could produce	2.67 ± 1.15	2	2	4
I believe advice from the system even when I don't know for certain that it is correct	2.67 ± 1.15	2	2	4
When I am uncertain about a decision I believe the system rather than myself	2.67 ± 0.58	2	3	3
Even if I have no reason to expect the system will be able to solve a difficult problem, I still feel certain that it will	2.33 ± 0.58	2	2	3
When the system gives unusual advice I am confident that the advice is correct	1.67 ± 0.58	2	2	1

System Trustworthiness Scale (items 1-9)

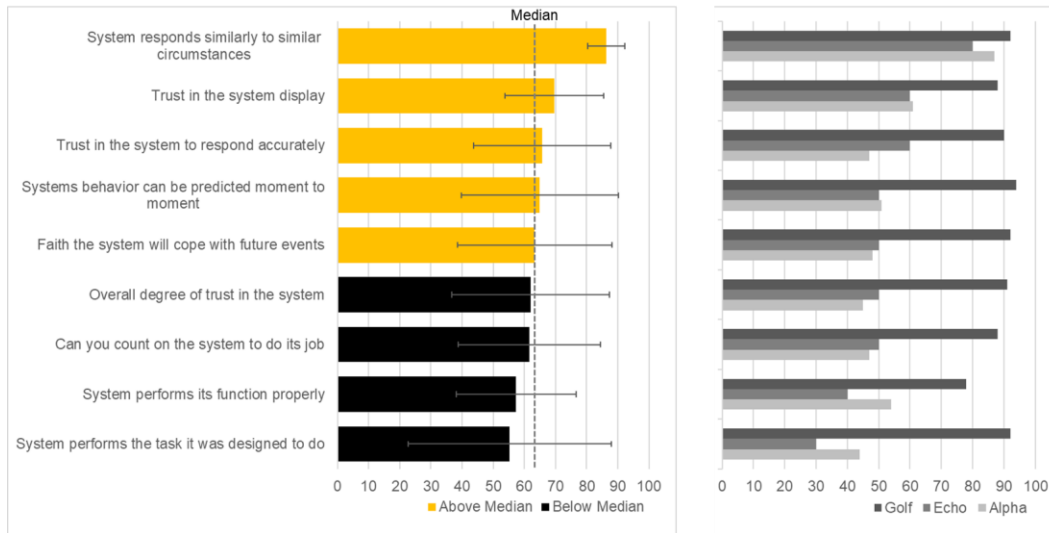


Figure A-7. Session 2 system trustworthiness scale.

Table A-7. Session 2 system trustworthiness scale.

Item query	Aggregate ($\mu \pm \sigma$)	Alpha	Echo	Golf
To what extent does the system respond similarly when it encounters similar circumstances at different points in time	86.33 $\pm$ 6.03	87	80	92
Your degree of trust in the system display	69.67 $\pm$ 15.89	61	60	88
Your degree of trust in the system to respond accurately	65.67 $\pm$ 22.05	47	60	90
To what extent can the systems behavior be predicted from moment to moment	65.00 $\pm$ 25.12	51	50	94
Your degree of faith that the system will be able to cope with future events	63.33 $\pm$ 24.85	48	50	92
Your overall degree of trust in the system	62.00 $\pm$ 25.24	45	50	91
To what extent can you count on the system to do its job	61.67 $\pm$ 22.85	47	50	88
To what extent does the system perform its function properly	57.33 $\pm$ 19.22	54	40	78
To what extent does the system perform the task it was designed to do	55.33 $\pm$ 32.52	44	30	92

SC3 Touchpoint Session 2 ITS Evaluation (item 3)

Did you manually adjust the tasking or properties of specific units within a given seed COA? If yes how many times (estimate), and why? [session 2]

Alpha:

“No”

Echo:

“Yes 10, I had to rearrange how the cav was tasked as they did not have a designated directed state insided the COA, Attempted to redefine how the maneuver units would be executing their tasks. Tasked the field artillery units to move to their assigned PAAs for the initial part of the mission. Identified that the unit that we lost the most 1-6 would need to be slowly moved into place as it was directly aligned to moving to monroe without conditions being set in madison.”

Golf:

“Yes 10 total units... 50+ total movements/task changes”

SC3 Touchpoint Session 2 ITS Evaluation (item 5)

Did you or your teammates verify any of COA-GPT recommendations independently? If so, what prompted that decision? [session 2]

Alpha:

“N/A”

Echo:

“We had to adjust the COA’s in some capacity to ensure that the entire mission through OBJ Madison could be accomplished. The initial set and directed actions seemed to be limited to seizing Madison and fixing Monroe as two sepearate actions that could be executed without concern related to time.”

Golf:

“No”

SC3 Touchpoint Session 3 ITS Evaluation (item 2)

Did you verify any of COA-GPT’s recommendations independently? What prompted that decision? [session 3]

Alpha:

“I verified a few of the tasks as it repeated the same task in multiple phases when it should have moved to the subsequent task.”

Echo:

“I went back to check starting locations that COA-GPT had recommended and attempted to update those locations related to positioning by phase. In some cases COA-GPT made those changes and some cases it did not even though it told me that it did.”

Golf:

“No”

SC3 Touchpoint Session 3 ITS Evaluation (item 3)

Please briefly describe whether you felt confident in the changes to the seed COAs that you made using COA-GPT, and why. [session 3]

Alpha:

“I felt somewhat confident in the changes I made to the seed COA in COA-GPT. Initially COA GPT updated my adjustments to 1-1CAV and 5-20IN in the sync matrix; however, when i made updates to 1-6IN it did not successfully translate them onto the sync matrix and adjusted 5-20IN to tasks and purposes not originally assigned.”

Echo:

“The changes that I made to the seed coas functioned the initial set as the exact changes made was reflected appropriately. For instance when I started the change I ended up making resulted in COA-GPT deleting out all the phases beside phase 1. This was translated over to the simulator and successfully represented that phase. However when I asked it to implement the rest of the phases again then portions of the phases returned but only in small aspects with 5-20 no longer having direction or purpose.”

Golf:

“The seed “COA” was not good. Once I went in and changed the COA to a better COA, the model did a good job portraying what I had written. It was not perfect, but it was a good start. It’s certainly worth continuing to work it.”

SC3 Touchpoint Session 3 ITS Evaluation (item 6 (a-c))

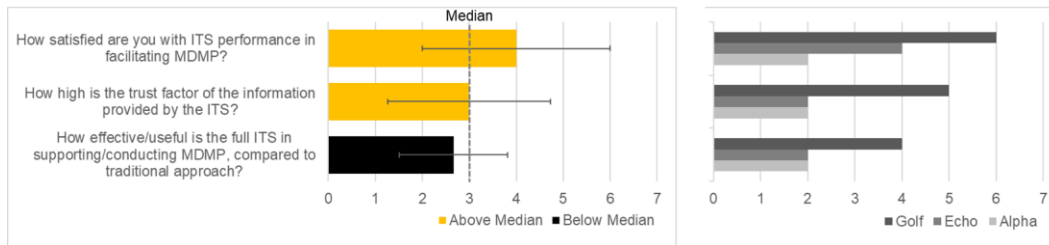


Figure A-8. Session 3 ITS evaluation.

Table A-8. Session 3 ITS evaluation.

Item query	Aggregate ($\mu \pm \sigma$)	Alpha	Echo	Golf
How satisfied are you with the ITSs performance in facilitating MDMP?	4.00 $\pm$ 2.00	2	4	6
How high is the trust factor of the information provided by the ITS?	3.00 $\pm$ 1.73	2	2	5
How effective and useful is the full ITS in supporting/conducting MDMP, compared to a traditional approach?	2.67 $\pm$ 1.15	2	2	4

SC3 Touchpoint Session 3 ITS Evaluation (item 1 (a-k))

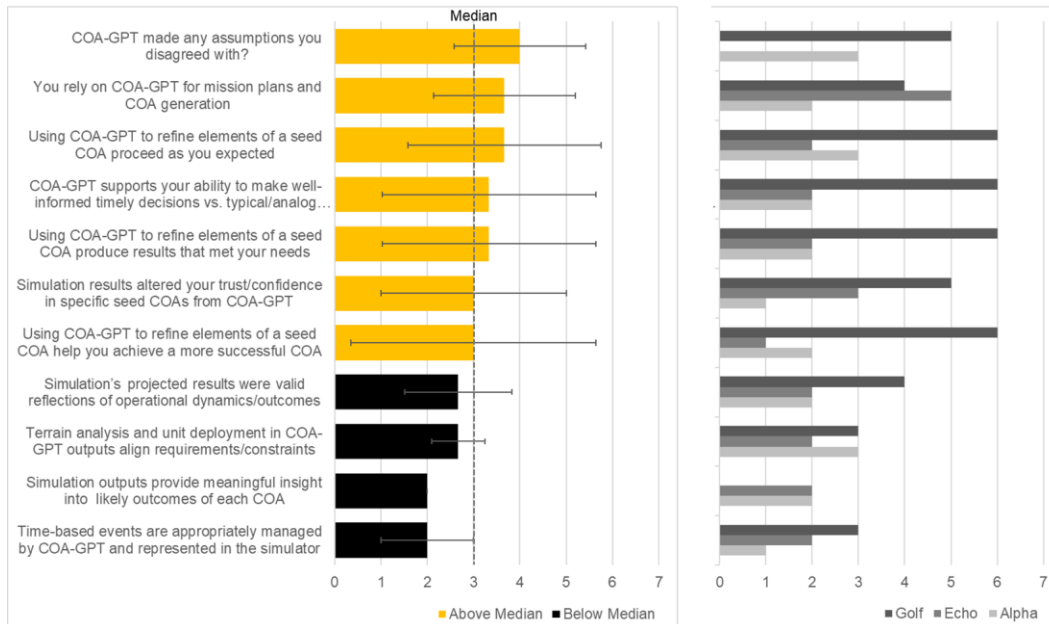


Figure A-9. Session 3 COA-GPT & simulator evaluation.

Table A-9. Session 3 COA-GPT & simulator evaluation.

Item query	Aggregate ($\mu \pm \sigma$)	Alpha	Echo	Golf
To what extent did you feel that COA-GPT made any assumptions you disagreed with?	4.00 $\pm$ 1.41	3	0	5
Did using COA-GPT to refine elements of a given seed COA proceed as you expected	3.67 $\pm$ 2.08	3	2	6
To what extent would you rely on COA GPT for mission plans and COA generation?	3.67 $\pm$ 1.53	2	5	4
To what extent did using COA-GPT to refine elements of a given seed COA produce results that met your needs	3.33 $\pm$ 2.31	2	2	6
To what extent did COA-GPT exchanges/requests support your ability to make well-informed timely decisions, relative to instances where you might perform these tasks in a typical, equivalent mission context (using analog tools or present-day technologies, e.g., CPCE)?	3.33 $\pm$ 2.31	2	2	6
To what extent did using COA-GPT to refine elements of a given seed COA help you achieve a more successful COA	3.00 $\pm$ 2.65	2	1	6
To what extent did results from the constructive simulation alter your trust or confidence in specific seed COAs generated by COA-GPT?	3.00 $\pm$ 2.00	1	3	5
To what extent did you feel that the simulation's projected results were valid reflections of operational dynamics and/or possible outcomes	2.67 $\pm$ 1.15	2	2	4
To what extent did you believe that the terrain analysis and unit deployment within COA-GPT outputs (as visualized/displayed in the simulator) to align with mission requirements and constraints	2.67 $\pm$ 0.58	3	2	3
To what extent were time-based events and phases appropriately managed by COA-GPT and represented in the simulator (e.g., operation start, phase shifts, timing of unit actions)?	2.00 $\pm$ 1.00	1	2	3
To what extent did the simulation outputs provide meaningful insight into the likely outcomes of each COA?	2.00 $\pm$ 0.00	2	2	0

Human-Computer Trust (short) (“COA-GPT” version)

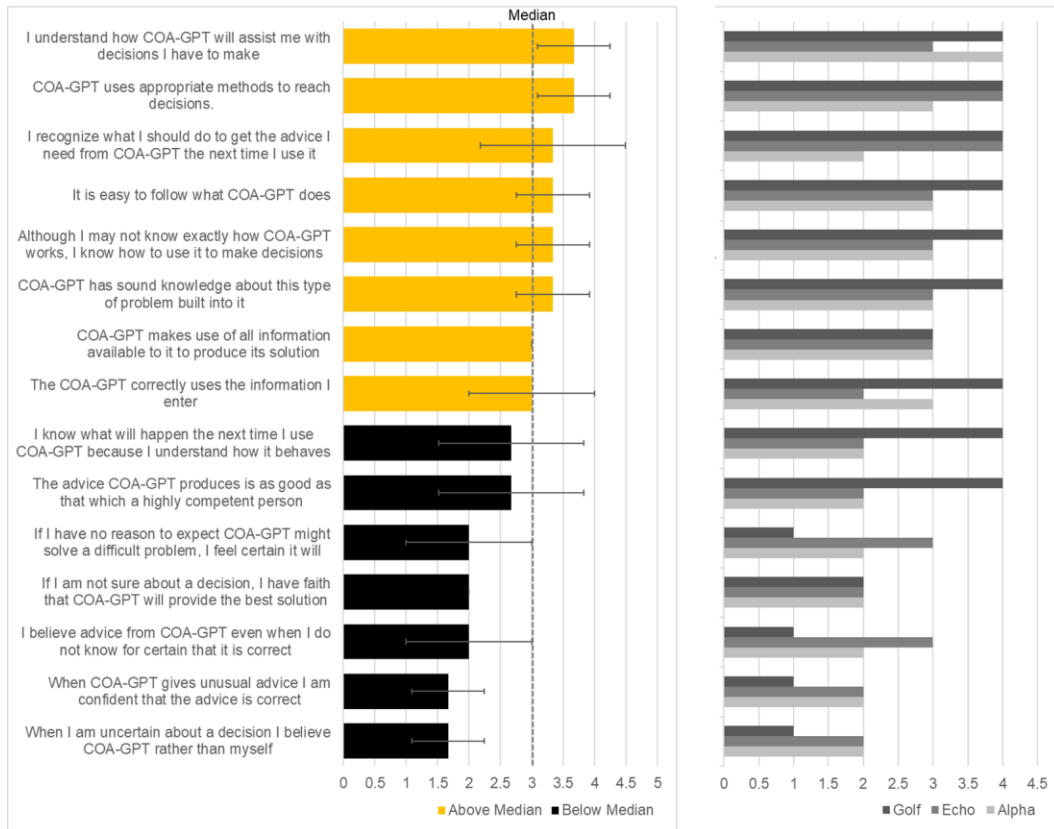


Figure A-10. Session 3 human-computer trust (COA-GPT).

Table A-10. Session 3 human-computer trust (COA-GPT).

Item query	Aggregate ($\mu \pm \sigma$)	Alpha	Echo	Golf
The COA-GPT component of the system uses appropriate methods to reach decisions.	3.67 ± 0.58	3	4	4
I understand how COA-GPT will assist me with decisions I have to make	3.67 ± 0.58	4	3	4
I recognize what I should do to get the advice I need from COA-GPT the next time I use it	3.33 ± 1.15	2	4	4
The COA-GPT component of the system has sound knowledge about this type of problem built into it	3.33 ± 0.58	3	3	4
Although I may not know exactly how COA-GPT works, I know how to use it to make decisions about the problem	3.33 ± 0.58	3	3	4
It is easy to follow what COA-GPT does	3.33 ± 0.58	3	3	4
The COA-GPT component of the system correctly uses the information I enter	3.00 ± 1.00	3	2	4
The COA-GPT component of the system makes use of all the knowledge and information available to it to produce its solution to the problem	3.00 ± 0.00	3	3	3
The advice that COA-GPT specifically produces is as good as that which a highly competent person could produce	2.67 ± 1.15	2	2	4
I know what will happen the next time I use COA-GPT because I understand how it behaves	2.67 ± 1.15	2	2	4
I believe advice from COA-GPT even when I don't know for certain that it is correct	2.00 ± 1.00	2	3	1
Even if I have no reason to expect COA-GPT will be able to solve a difficult problem, I still feel certain that it will	2.00 ± 1.00	2	3	1
If I am not sure about a decision, I have faith that COA-GPT will provide the best solution	2.00 ± 0.00	2	2	2
When I am uncertain about a decision I believe COA-GPT rather than myself	1.67 ± 0.58	2	2	1
When COA-GPT gives unusual advice I am confident that the advice is correct	1.67 ± 0.58	2	2	1

A.1.2 Team Performance

SC3 Touchpoint Session 2 ITS Evaluation (item 6)

*Did ITS outputs prompt any disagreements or discussions among team members?
If so, what were they about? [session 2]*

Alpha:

“Not all of the tasks from COA GPT were reflected in the ITS.”

Echo:

“We used three of the four different COAs to differentiate and then tried to find the best way to allow for the largest amount of combat power available after fixing OBJ Monroe. All three of us went about trying the tool separately from reorganizing the taskings of our maneuver units, field artillery units, and screening units to include the cav. One member found that giving directed movement actions and then destroy tasks directly to units provided the best results while I was more attempting to get the simulation to fight for itself without me having to give that level of direction.”

Golf:

“No”

SC3 Touchpoint Session 2 ITS Evaluation (item 7)

How did the use of the technologies influence the sequencing and timing of your planning activities compared to when you operated the ITS solo? [session 2]

Alpha:

“We set time hacks to have discussion and had more interactions about what we were seeing with our different COAs and how we were tasking them.”

Echo:

“We got together as a group to see what the initial four COAs provided as results. Once we determined the ones that for whatever reason provided the same combat power results we then separated out to use the other three COAs to manipulate and try to improve blue’s combat power. While we used the configuration of units and initial tasks the COA description did not always align to the actual execution of the simulation as they all seemed to take the same actions between three of the four. The simulation needed more detailed direction to allow for dynamic shifts in the protection of the blue forces. The red forces by default were never given any other tasks but to be in place. This would have to be something that would then have to be run by an OPFOR to give a more adequate representation of what the enemy is going to do in this action.”

Golf:

“Currently, sequencing of events that must occur are not done well...

Example: Clearing PAAs must be done prior to seizing Obj MADISON.”

SC3 Touchpoint Session 2 ITS Evaluation (item 9)

In a future mission, would you prefer to use ITS individually or as a team? Why? [session 2]

Alpha:

“I would prefer it as a team. This allows the different Warfighting function subject matter experts the ability to collaborate and better develop the COA.”

Echo:

“If I could use it as a team with the ability to manipulate certain parts of the COA while identifying the change in some way that would be great.”

Golf:

“Once it’s learned/mastered, it would be useful with a team. However, at the current time it’s good for use individually.”

SC3 Touchpoint Session 2 ITS Evaluation (item 10)

Would you advise integrating ITS into formal team structures or command roles in future missions? Why or why not? [session 2]

Alpha:

“With more refinement and updates I would recommend the ITS for formal team structures in the future. It has the potential to quickly review multiple COAs and analyze it with given COA Evaluation Criteria. This would help speed up the overall process.”

Echo:

“I think the development of the ITS should be done with multiple players, Warfighting functions participating to ensure that we all believe it viable. Initially the idea would have to be implemented with different users playing as their roles with one version of the simulation, not sure if it would be able to be broken down to combine in future versions.”

Golf:

“I would integrate. The system is currently in a nascent state. As it’s worked on, the system will become far better. We saw how far COA-GPT progressed in just a few months. I’m optimistic that this system will get better very quickly.”

SC3 Touchpoint Session 2 ITS Evaluation (item 11)

Please explain any reasoning for your opinions of the ITS as reflected above [session 2]

Alpha:

“While I do not have high confidence in the ITS at the moment at its current level. I do believe that it has great potential and with more refinement can be a critical tool to give commanders and staffs a significant advantage on the battlefield.”

Echo:

“Within a couple of hours of comfort with the tool I was able to determine that I could take a baseline breakdown of a task organization predefined by COA GPT and start to provide it tasks to execute the mission that I would want it to do. The simulation is still wanting in that it does not easily identify failure rates without some investigating on behalf of the user and backtracking. The additional actions of the user is to use the feedback for units in combat to then direct actions against those units to refine. I see that as a success to exploit as it then allows for gaming the system. There still remain limitations as the enemy has to have a plan of action validated by the intelligence or OPFOR team and in its current setting the units are static but have the ability to engage. Being able to further play with units that have accurate range ability and sequencing through a matrix for all the elements is necessary. In each COA the focus was on the maneuver units and supporting elements as well as protection elements were not in play in a meaningful way as of yet.

Survivability is key to the COA selection criteria.

Some Improvements:

The ability to copy tasks would be helpful.

How do we get the tool to define waiting to move from OBJ Madison to OBJ Monroe.

Tell me in space and time how long I assume to move will take.

When it tells me to that it will clear 7158, I don't know what that means, 7158 BN?

How do you mass Fires, direct fires on a target, consolidate to wait, etc.

Need to be able to determine who does and does not have a task.

Need a search function for units”

Golf:

“As stated above, as the system is updated/refined, the system will quickly become far better. We saw how far COA-GPT progressed in just a few months. I’m optimistic that this system will get better very quickly.”

SC3 Touchpoint Session 2 ITS Evaluation (item 4 (l-q only))

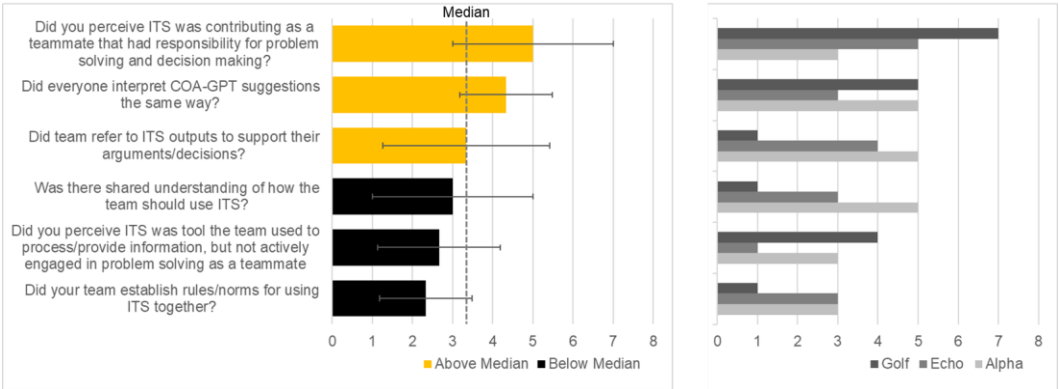


Figure A-11. Session 2 ITS evaluation - team.

Table A-11. Session 2 ITS evaluation - team.

Item query	Aggregate ($\mu \pm \sigma$)	Alpha	Echo	Golf
Did you perceive that COA-GPT and the simulator interface were contributing as teammates that had responsibility for problem solving and were engaged actively in the team's decision making process	5.00 $\pm$ 2.00	3	5	7
Did everyone interpret the COA-GPTs suggestions the same way?	4.33 $\pm$ 1.15	5	3	5
Did team members refer back to ITS outputs to support their arguments or decisions?	3.33 $\pm$ 2.08	5	4	1
Was there a shared understanding of how the team should use ITS during the session?	3.00 $\pm$ 2.00	5	3	1
Did you perceive that COA-GPT and the simulator interface were tools that you and your team used to process and provide information, but did not actively engage in problem solving or decision making as a teammate	2.67 $\pm$ 1.53	3	1	4
Did your team establish any rules or norms for using ITS together (e.g., who enters data, who interprets output)?	2.33 $\pm$ 1.15	3	3	1

SC3 Touchpoint Session 2 Team Process Scale

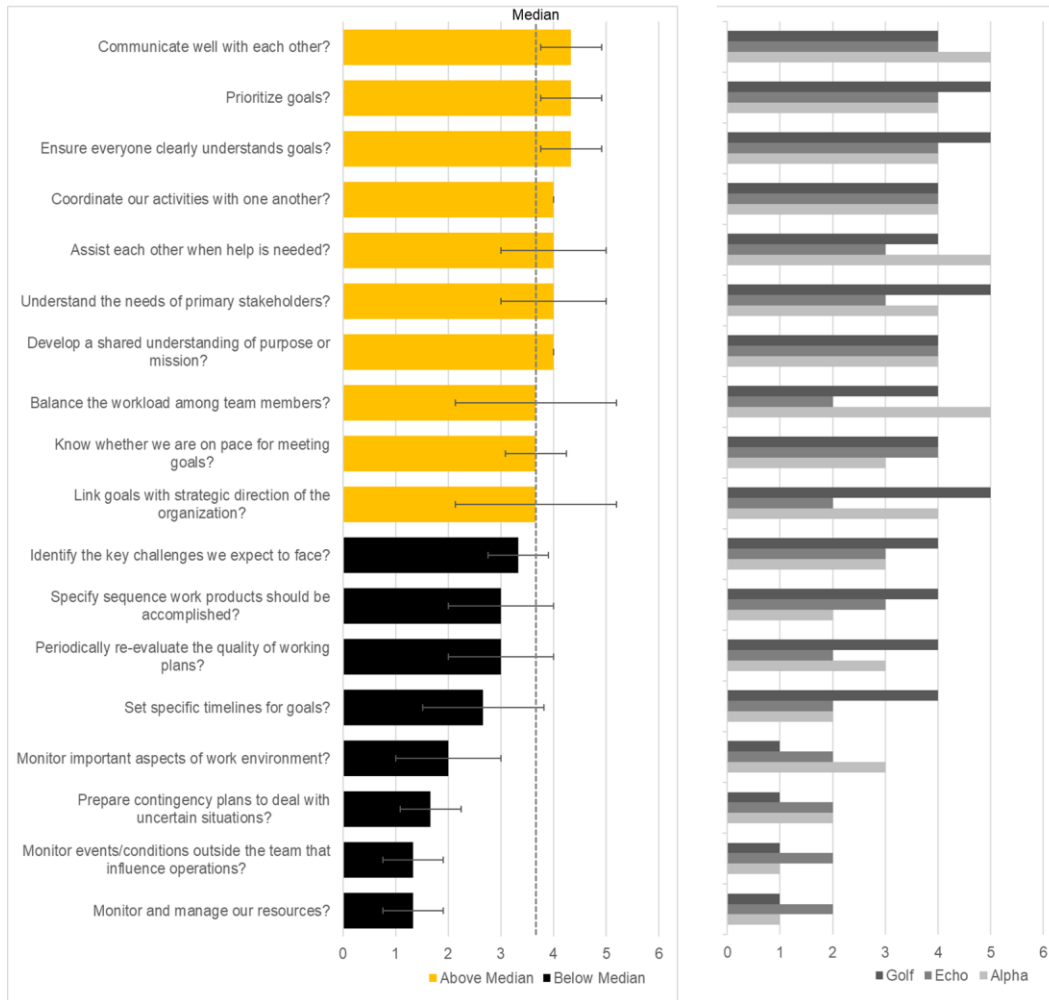


Figure A-12. Session 2 team process scale.

Table A-12. Session 2 team process scale.

Item query	Aggregate ($\mu \pm \sigma$)	Alpha	Echo	Golf
Ensure that everyone on our team clearly understands our goals?	4.33 ± 0.58	4	4	5
Prioritize our goals?	4.33 ± 0.58	4	4	5
Communicate well with each other?	4.33 ± 0.58	5	4	4
Understand the needs of our primary stakeholders (e.g., commander)?	4.00 ± 1.00	4	3	5
Assist each other when help is needed?	4.00 ± 1.00	5	3	4
Develop a shared understanding of our purpose or mission?	4.00 ± 0.00	4	4	4
Coordinate our activities with one another?	4.00 ± 0.00	4	4	4
Link our goals with the strategic direction of the organization?	3.67 ± 1.53	4	2	5
Balance the workload among our team members?	3.67 ± 1.53	5	2	4
Know whether we are on pace for meeting our goals?	3.67 ± 0.58	3	4	4
Identify the key challenges that we expect to face?	3.33 ± 0.58	3	3	4
Periodically re-evaluate the quality of our working plans?	3.00 ± 1.00	3	2	4
Specify the sequence in which work products should be accomplished?	3.00 ± 1.00	2	3	4
Set specific timelines for each of our goals?	2.67 ± 1.15	2	2	4
Monitor important aspects of our work environment (e.g., inventories, equipment and process operations, information flows)?	2.00 ± 1.00	3	2	1
Prepare contingency plans to deal with uncertain situations?	1.67 ± 0.58	2	2	1
Monitor events and conditions outside the team that influence our operations?	1.33 ± 0.58	1	2	1

A.1.3 Design Feedback

Questionnaire for User Interface Satisfaction

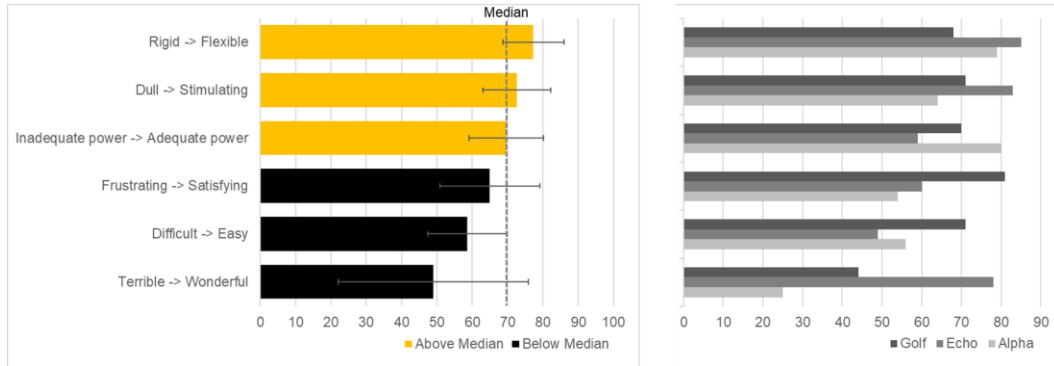


Figure A-13. Session 1 questionnaire for user interface satisfaction – overall (items 1-6).

Table A-13. Session 1 questionnaire for user interface satisfaction – overall (items 1-6).

Item query	Aggregate ($\mu \pm \sigma$)	Alpha	Echo	Golf
Inadequate power ↔ Adequate power	77.33 ± 8.62	79	85	68
Dull ↔ Stimulating	72.67 ± 9.61	64	83	71
Difficult ↔ Easy	69.67 ± 10.50	80	59	70
Terrible ↔ Wonderful	65.00 ± 14.18	54	60	81
Frustrating ↔ Satisfying	58.67 ± 11.24	56	49	71
Rigid ↔ Flexible	49.00 ± 26.85	25	78	44

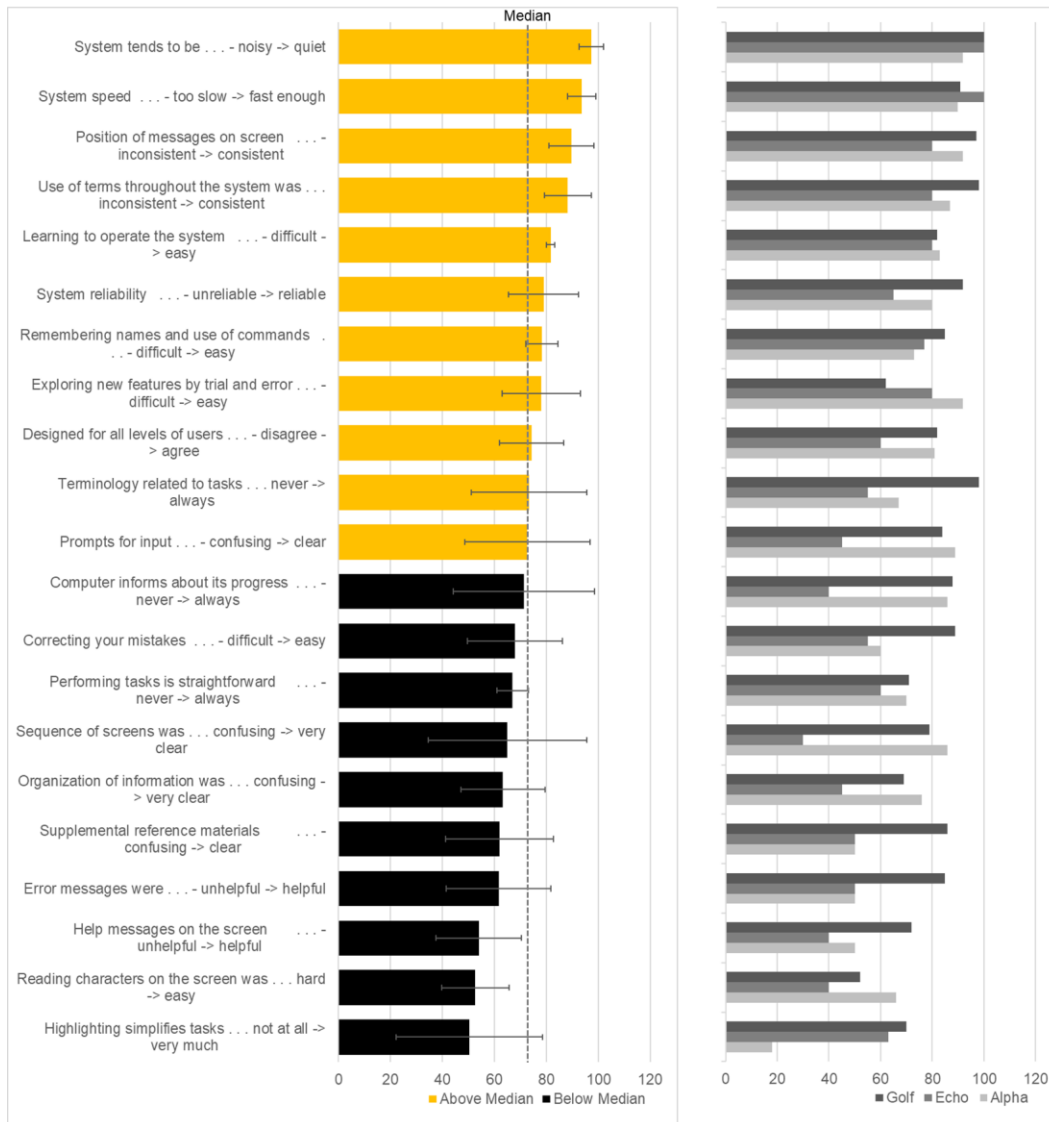


Figure A-14. Session 1 questionnaire for user interface satisfaction – features (items 7-28).

Table A-14. Session 1 questionnaire for user interface satisfaction – features (items 7-28).

Item query	Aggregate ($\mu \pm \sigma$)	Alpha	Echo	Golf
System tends to be: Noisy ↔ Quiet	97.33 ± 4.62	92	100	100
System speed: Too slow ↔ Fast enough	93.67 ± 5.51	90	100	91
Position of messages on screen: Inconsistent ↔ Consistent	89.67 ± 8.74	92	80	97
Use of terms throughout the system was: Inconsistent ↔ Consistent	88.33 ± 9.07	87	80	98
Learning to operate the system: Difficult ↔ Easy	81.67 ± 1.53	83	80	82
System reliability: Unreliable ↔ Reliable	79.00 ± 13.53	80	65	92
Remembering names and use of commands: Difficult ↔ Easy	78.33 ± 6.11	73	77	85
Exploring new features by trial and error: Difficult ↔ Easy	78.00 ± 15.10	92	80	62
Designed for all levels of users: Disagree ↔ Agree	74.33 ± 12.42	81	60	82
Terminology related to tasks: Never ↔ Always	73.33 ± 22.19	67	55	98
Prompts for input: Confusing ↔ Clear	72.67 ± 24.09	89	45	84
Computer informs about its progress: Never ↔ Always	71.33 ± 27.15	86	40	88
Correcting your mistakes: Difficult ↔ Easy	68.00 ± 18.36	60	55	89
Performing tasks is straightforward: Never ↔ Always	67.00 ± 6.08	70	60	71
Sequence of screens was: Confusing ↔ Clear	65.00 ± 30.51	86	30	79
Organization of information was: Confusing ↔ Clear	63.33 ± 16.26	76	45	69
Supplemental reference materials: Confusing ↔ Clear	62.00 ± 20.78	50	50	86
Error messages were: Unhelpful ↔ Helpful	61.67 ± 20.21	50	50	85
Help messages on the screen: Unhelpful ↔ Helpful	54.00 ± 16.37	50	40	72
Reading characters on the screen was: Hard ↔ Easy	52.67 ± 13.01	66	40	52
Highlighting simplifies tasks: Not at all ↔ Very much	50.33 ± 28.22	18	63	70

Technology Acceptance Model

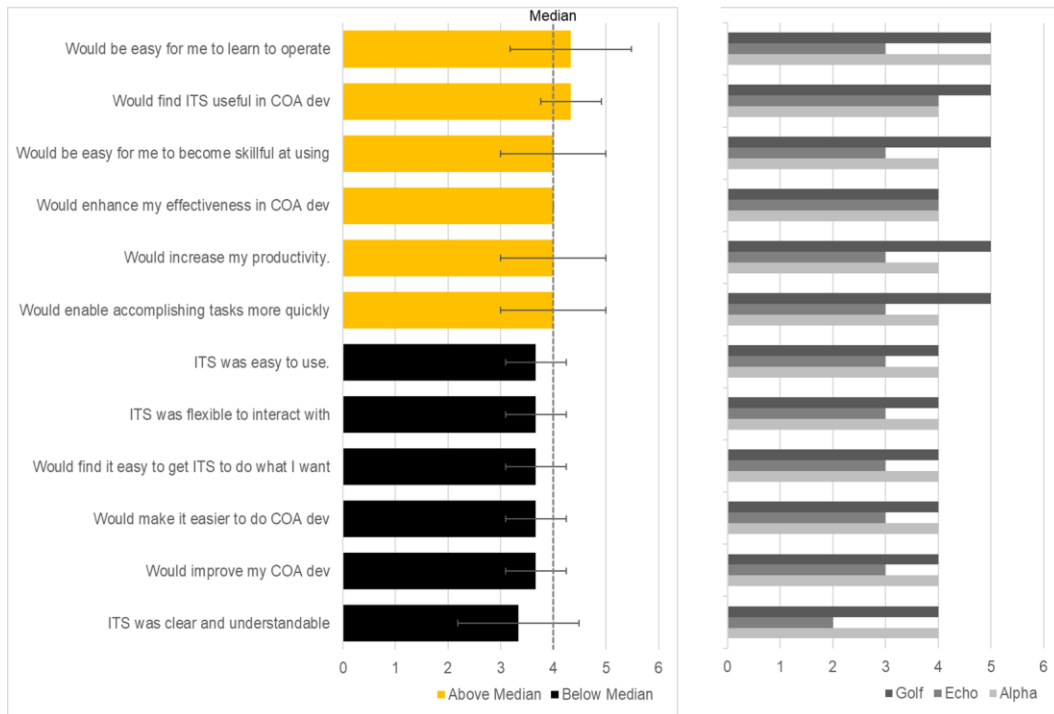


Figure A-15. Session 2 technology acceptance model.

Table A-15. Session 2 technology acceptance model.

Item query	Aggregate ($\mu \pm \sigma$)	Alpha	Echo	Golf
Learning to operate the ITS would be easy for me.	4.33 ± 1.15	5	3	5
I would find the ITS useful in COA development.	4.33 ± 0.58	4	4	5
Using the ITS in COA development would enable me to accomplish tasks more quickly	4.00 ± 1.00	4	3	5
Using the ITS in COA development would increase my productivity.	4.00 ± 1.00	4	3	5
It would be easy for me to become skillful at using the ITS.	4.00 ± 1.00	4	3	5
Using the ITS would enhance my effectiveness in COA development.	4.00 ± 0.00	4	4	4
Using the ITS would improve my COA development performance.	3.67 ± 0.58	4	3	4
Using the ITS would make it easier to do COA development	3.67 ± 0.58	4	3	4
I would find it easy to get the ITS to do what I want to do	3.67 ± 0.58	4	3	4
I find the ITS flexible to interact with.	3.67 ± 0.58	4	3	4
I find the ITS easy to use.	3.67 ± 0.58	4	3	4
My interaction with the ITS was clear and understandable	3.33 ± 1.15	4	2	4

Design Features Feedback

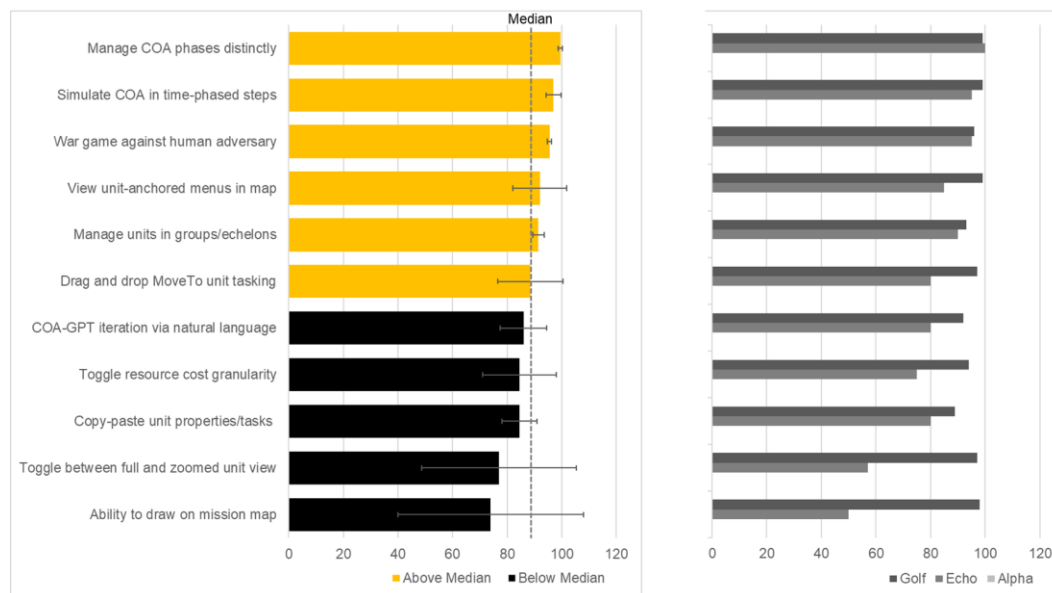


Figure A-16. Session 2 design feature ratings. Responses for SME Alpha were not captured for this questionnaire.

Table A-16. Session 2 design feature ratings.

Item query	Aggregate ($\mu \pm \sigma$)	Alpha	Echo	Golf
The ability to manage COA phases distinctly such that one could modify the plans for units or sets of units up to a certain time or phase line, and then proceed to adjusting plans for the next phase	99.50 $\pm$ 0.71	N/A	100	99
The ability to run the simulation of the COA in phases or up to certain time points to allow you to inspect when/where events occur and issues arise	97.00 $\pm$ 2.83	N/A	95	99
The ability to war game against another human who is tasked with developing a COA for the enemy forces (note: this capability would perhaps depend on being able to run the simulation and re-plan in phases or at critical decision points)	95.50 $\pm$ 0.71	N/A	95	96
The ability to view unit anchored menus within the map view that allow you to specify unit properties and tasks within that view rather than using the taskbar and drop-down menus (that is, the UI tools on the left)	92.00 $\pm$ 9.90	N/A	85	99
The ability to manage units at different echelons and to designate/create groups of units that may be given the same sets of tasks all at once rather than having to set those tasks individually for each unit	91.50 $\pm$ 2.12	N/A	90	93
The ability to set unit positions for MoveTo (i.e., movement) tasks by dragging and dropping units and designating the time (or mission phase) by which they should reach the given destination	88.50 $\pm$ 12.02	N/A	80	97
The ability to use COA-GPT to engage in additional, natural language iterations that allow you to modify specific components of a given seed COA (e.g., to use natural language to alter the COA for a specific unit such as C 1-1 or for a given phase of the mission rather than modifying their plan manually through the interface)	86.00 $\pm$ 8.49	N/A	80	92
The ability to copy-paste properties and/or tasks from one unit (or set of units) to another	84.50 $\pm$ 6.36	N/A	80	89
The ability to toggle granularity of resource costs in the simulator outputs to allow you to drill down into COA costs	84.50 $\pm$ 13.44	N/A	75	94
The ability to toggle back and forth between the full overview of the operational area and a zoomed in view such that one may be zoomed into and altering the tasking for a given unit, toggle to zoom out to the full view to review positions, and then toggle again to return to the unit one was inspecting	77.00 $\pm$ 28.28	N/A	57	97
The ability to draw on the mission map to designate important geometries such as regions that are the “domain” of one or another unit or perhaps represent avenues of approach, etc.	74.00 $\pm$ 33.94	N/A	50	98

Notes: Responses for SME Alpha were not captured for this questionnaire.

A.2. Self-Report Instruments

The following section catalogs the self-report instruments administered during the touchpoint.

A.2.1 User Traits and Individual Differences

A.2.1.1 Demographics

Administration Notes

Sources/references

The items in this survey were developed partly in service of the SC3 Touchpoint.

Notes

The response options for this survey were varied and are embedded in the questions below for clarity.

Measures

- 1) What is your rank? [response options: free response]
- 2) What is your job/classification? [response options: free response]
- 3) Have you participated in COA development as part of your duties [response items: yes, no]?
 - a. If so, how many years experience do you have with COA development? [response options: free response]
- 4) What is your level of experience with machine learning and artificial intelligence? [response options: None, Familiar with terms and concepts, Experiences (some understanding of ML methods, Expert (work with AI/ML in robotics), Expert (work with AI/ML in other applications)]
- 5) Have you interacted with COA-GPT or the simulator interface prior to this touchpoint?
- 6) Have you worked with the Tropic Tortoise scenario prior to this touchpoint?

A.2.1.2 Rational and Intuitive Decision Style Scale

Administration Notes

The Rational and Intuitive Decision styles scale is a 10-item scale that assesses the rational and intuitive dimensions of an individual's preferred decision style. Users rate each item using 5-point Likert scales (1 – Strongly Disagree to 5 - Strongly Agree). There is a high test-retest reliability for both the rational ($r = .79$, $p < .01$) and intuitive ($r = .79$, $p < .01$) dimensions.

Sources/references

Hamilton, K., Shih, S. I., & Mohammed, S. (2016). The development and validation of the rational and intuitive decision styles scale. *Journal of personality assessment*, 98(5), 523-535. DOI: 10.1080/00223891.2015.1132426.

Notes

All items employed a 5-point Likert scale (response options: strongly disagree, disagree, neutral, agree, strongly agree).

Measures

The following questions relate to how you make decisions. There are no “right” or “wrong” answers, so please state your opinion as honestly as possible. Using the scale below, please indicate the extent to which you agree or disagree with the statements. Describe how you are now, not as you wish to be in the future.

Rational Items:

- 1) I prefer to gather all the necessary information before committing to a decision.
- 2) I thoroughly evaluate decision alternatives before making a final choice.
- 3) In decision making, I take time to contemplate the pros/cons or risks/benefits of a situation.
- 4) Investigating the facts is an important part of my decision-making process.
- 5) I weigh a number of different factors when making decisions

Intuitive Items:

- 6) When making decisions, I rely mainly on my gut feelings.
- 7) My initial hunch about decisions is generally what I follow.
- 8) I make decisions based on intuition.
- 9) I rely on my first impressions when making decisions.
- 10) I weigh feelings more than analysis in making decisions.

A.2.1.3 Intolerance of Uncertainty Scale

Administration Notes

The Intolerance of Uncertainty Scale – Short Format (Buhr & Dugas, 2002; Carelton, Norton, & Asmundson, 2007) assesses the tendency of an individual to respond inappropriately to the possibility that a negative event may occur. The shortened survey is a 12-item questionnaire that assesses two factors, prospective anxiety and inhibitory anxiety, using 54-point Likert scales (Appendix H). Each factor has high internal reliability, prospective anxiety 7 items, $\alpha = .85$ and

inhibitory anxiety, 5 items, $\alpha = .85$). The shortened version highly correlates with the original 27 item survey ($r = .96$).

Sources/references

Buhr, K., & Dugas, M. J. (2002). The intolerance of uncertainty scale: Psychometric properties of the English version. *Behaviour research and therapy*, 40(8), 931-945.

Carleton, R. N., Norton, M. P. J., & Asmundson, G. J. (2007). Fearing the unknown: A short version of the Intolerance of Uncertainty Scale. *Journal of anxiety disorders*, 21(1), 105-117.

Notes

All items employed a 5-point Likert scale (response options: Not at all, Somewhat, Neutral, Very, Entirely).

Measures

Please respond to the prompts below in terms of whether each statement is characteristic of you.

Prospective Anxiety:

- 1) Unforeseen events upset me greatly
- 2) It frustrates me not having all the information I need
- 3) One should always look ahead so as to avoid surprises
- 4) A small, unforeseen event can spoil everything, even with the best of planning
- 5) I always want to know what the future has in store for me
- 6) I can't stand being taken by surprise
- 7) I should be able to organize everything in advance

Inhibitory Anxiety:

- 8) Uncertainty keeps me from living a full life
- 9) When it's time to act, uncertainty paralyses me
- 10) When I am uncertain I can't function very well
- 11) The smallest doubt can stop me from acting.
- 12) I must get away from all uncertain situations.

A.2.1.4 Propensity to Trust (LLM Agents)

Administration Notes

Sources/references

Jessup, S. A., Schneider, T. R., Alarcon, G. M., Ryan, T. J., & Capiola, A. (2019, July). The measurement of the propensity to trust automation. In International Conference on Human-Computer Interaction (pp. 476-489). Springer, Cham.

Schneider, T. R., Jessup, S. A., Stokes, C., Rivers, S., Lohani, M., & McCoy, M. (2017, May). The influence of trust propensity on behavioral trust. Poster session presented at the meeting of Association for Psychological Society, Boston, MA.

Notes

All items employed a 5-point Likert scale (response options: Strongly disagree, Disagree, Neutral, Agree, Strongly agree).

Measures

For the below listed items, please read each statement carefully. Using the 5-point scale ranging from 1 (strongly disagree) to 5 (strongly agree), select the answer that most accurately describes your feelings.

- 1) Generally, I trust LLM agents.
- 2) LLM agents help me solve many problems.
- 3) I think it's a good idea to rely on LLM agents for help.
- 4) I don't trust the information I get from LLM agents.
- 5) LLM agents are reliable.
- 6) I rely on LLM agents

A.2.1.5 Team Process Scale (Short)

Administration Notes

The Team Process Scale (Mathieu et al 2020) is a 50-item questionnaire that assesses team processes by evaluating 10 first level constructs that group onto 3 higher-order constructs (i.e., transition processes, action processes, interpersonal processes). The individual scales demonstrated high internal consistency ($\alpha > .90$). The questionnaire can be customized to focus on those constructs of specific interest to the researcher and appropriate for the setting. In this study, we have reduced the items to 18, assessing two higher-order constructs; transition and action processes. Each item is scored using 5-point Likert scales, where 1 = not at all, and 5 = to a very great extent. Higher scores indicate better functioning team processes, lower scores indicate areas that could be improved.

Sources/references

Mathieu, J. E., Luciano, M. M., D’Innocenzo, L., Klock, E. A., & LePine, J. A. (2020). The development and construct validity of a team processes survey measure. *Organizational Research Methods*, 23(3), 399–431. <https://doi.org/10.1177/1094428119840801>.

Notes

All items employed a 5-point Likert scale (response options: Strongly disagree, Disagree, Neutral, Agree, Strongly agree).

Measures

To what extent did our team actively work to ...

Transition Processes:

Mission Analysis:

- 1) Identify the key challenges that we expect to face?
- 2) Develop a shared understanding of our purpose or mission?
- 3) Understand the needs of our primary stakeholders (e.g., commander)?

Goal Specification:

- 4) Ensure that everyone on our team clearly understands our goals?
- 5) Link our goals with the strategic direction of the organization?
- 6) Prioritize our goals?
- 7) Set specific timelines for each of our goals?

Strategy Formation and Planning:

- 8) Prepare contingency (“if-then”) plans to deal with uncertain situations?
- 9) Periodically re-evaluate the quality of our working plans?
- 10) Specify the sequence in which work products should be accomplished?

Action Processes:

Monitoring Progress towards Goals:

- 11) Know whether we are on pace for meeting our goals?

Systems Monitoring:

- 12) Monitor and manage our resources (e.g., available units, ammo, etc.)?
- 13) Monitor important aspects of our work environment (e.g., inventories, equipment and process operations, information flows)?
- 14) Monitor events and conditions outside the team that influence our operations?

Team Monitoring and Backup:

- 15) Balance the workload among our team members?
- 16) Assist each other when help is needed?

Coordination:

- 17) Communicate well with each other?
- 18) Coordinate our activities with one another?

A.2.2 Workload

A.2.2.1 NASA-TLX (Modified)

Administration Notes

The NASA-TLX is a self-report questionnaire of perceived demands in six areas: mental, physical, temporal, effort (mental and physical), frustration, and performance. An Overall Workload rating will be obtained using the mean of the individual factor scores. Each factor has been shown to contribute independent information regarding the specific source of workload within a task, as such, in addition to an overall workload rating we will evaluate separately the contribution of each factor.

Sources/references

Hart, Sandra G. (October 2006). "NASA-Task Load Index (NASA-TLX); 20 Years Later" (PDF). Proceedings of the Human Factors and Ergonomics Society Annual Meeting. 50 (9): 904–908

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. *Advances in psychology*, 52, 139-183.

Notes

All items employed a response scale from Low->High (responses provided using a slider 0->100).

Measures

Please rate your overall impression of the demands imposed on you working with the ITS (combined COA-GPT and simulator interface) during the exercise with respect to whether the demand was low or high.

- 1) Mental Demand: How much mental and perceptual activity was required (e.g., thinking, looking, searching, etc.)? Was the task easy or demanding, simple or complex, exacting or forgiving?
- 2) Physical Demand: How much physical activity was required (e.g., pushing, pulling, turning, controlling, activating, etc.)? Was the task easy or demanding, slow or brisk, slack or strenuous, restful or laborious?
- 3) Temporal Demand: How much time pressure did you feel due to the rate or pace at which the task or task elements occurred? Was the pace slow and leisurely or rapid and frantic?
- 4) Level of Effort: How hard did you have to work (mentally and physically) to accomplish your level of performance?
- 5) Level of Frustration: How insecure, discouraged, irritated, stressed and annoyed versus secure, gratified, content, relaxed and complacent did you feel during the task?
- 6) Performance demand: How successful do you think you were in accomplishing the goals of the task set by the experimenter (or yourself)? How satisfied were you with your performance in accomplishing these goals?

A.2.2.2 Team Workload Questionnaire

Administration Notes

The Team Workload Questionnaire (TWLQ) was developed identifying three separate components of team workload, with workload in teams consisting of Task Workload, Team Workload and Task-Team Balancing. The Team Workload ($\alpha = .74$) aspect of the questionnaire is concerned with the demands of team interactions, specifically communicating, coordinating and the monitoring of team performance. In this study only the Team Workload portion of the questionnaire was administered.

Sources/references

Sellers, J., Helton, W. S., Näswall, K., Funke, G. J., & Knott, B. A. (2014, September). Development of the team workload questionnaire (TWLQ). In Proceedings of the human factors and ergonomics society annual meeting (Vol. 58, No. 1, pp. 989-993). Sage CA: Los Angeles, CA: Sage Publications.

Notes

The following questions relate to the team's performance in the task, you should answer them by considering the team as a whole. All items employed a response scale from Low->High (responses provided using a slider 0->10).

Measures

Please rate your overall impression of demands imposed on you during the exercise.

- 1) Coordination Demand: How much coordination activity was required (e.g., correction, adjustment)? Were the coordination demands to work as a team low or high, infrequent or frequent?
- 2) Communication Demand: How much communication activity was required (e.g., discussing, negotiating, sending and receiving messages)? Were the communication demands low or high, infrequent or frequent, simple or complex?
- 3) Time Sharing Demand: How difficult was it to share and manage time between taskwork (work done individually) and teamwork (work done as a team)? Was it easy or hard to manage individual tasks and those tasks requiring work with other team members?
- 4) Team Effectiveness: How successful do you think the team was in working as a team? How satisfied were you with the team-related aspects of performance?
- 5) Team Support: How difficult was it to provide and receive support (providing guidance, helping team members, providing instructions, etc.) from team members? Was it easy or hard to support/guide and receive support/guidance from other team members?
- 6) Team Dissatisfaction: How emotionally draining and irritating versus emotionally rewarding and satisfying was it to work as a team?

A.2.3 System Evaluations

A.2.3.1 Human-Computer Trust (Short)

Administration Notes

Human-Computer Trust (HCT) Questionnaire (Madsen & Gregor, 2000): The HCT questionnaire assesses five constructs of trust in intelligent decision aids and encompasses both affective and cognitive based trust. Each construct has high internal consistency, construct validity, and reliability (i.e., Perceived Reliability $\alpha = 0.85$, Perceived Technical Competence $\alpha = 0.74$, Perceived Understandability $\alpha = 0.84$, Faith $\alpha = 0.88$, and Personal Attachment $\alpha = 0.90$). Each construct evaluation consists of 5 items scored on Likert scales. Scores can be averaged or summed; higher scores indicate more positive associations.

The “COA-GPT” version of this instrument was administered during Session 3, wherein the prompt specifically requested responses with respect to “COA-GPT” only, rather than the ITS as a whole.

Sources/references

Madsen, M., & Gregor, S. (2000, December). Measuring human-computer trust. In 11th australasian conference on information systems (Vol. 53, pp. 6-8).

Notes

All items employed a 5-point Likert scale (response options: strongly disagree, disagree, neutral, agree, strongly agree).

Measures

Please respond to the following questions regarding your experience with the Information Technology System (e.g., the combination of the COA-GPT and Simulator Interface).

Perceived Technical Competence:

- 1) The system uses appropriate methods to reach decisions.
- 2) The system has sound knowledge about this type of problem built into it
- 3) The advice the system produces is as good as that which a highly competent person could produce
- 4) The system correctly uses the information I enter
- 5) The system makes use of all the knowledge and information available to it to produce its solution to the problem

Perceived Understandability:

- 6) I know what will happen the next time I use the system because I understand how it behaves
- 7) I understand how the system will assist me with decisions I have to make
- 8) Although I may not know exactly how the system works, I know how to use it to make decisions about the problem
- 9) It is easy to follow what the system does
- 10) I recognize what I should do to get the advice I need from the system the next time I use it.

Faith:

- 11) I believe advice from the system even when I don't know for certain that it is correct
- 12) When I am uncertain about a decision I believe the system rather than myself
- 13) If I am not sure about a decision, I have faith that the system will provide the best solution
- 14) When the system gives unusual advice I am confident that the advice is correct.

- 15) Even if I have no reason to expect the system will be able to solve a difficult problem, I still feel certain that it will.

A.2.3.2 User Interface Satisfaction

Administration Notes

Sources/references

Chin, J. P., Diehl, V. A, Norman, K. (Sept. 1987) Development of an instrument measuring user satisfaction of the human-computer interface, Proc. ACM CHI '88 (Washington, DC) 213-218. CS-TR-1926, CAR-TR-328

Harper, P. & Norman, Kent. (1993). Improving user satisfaction: The questionnaire for user interaction satisfaction version 5.5.

Notes

Items 1 to 5 employed a response scale from Low to High (responses provided using a slider 0->100). Higher scores indicate more positive evaluations. Items 6 and 7 employed free response.

Measures

Please use the scale below to describe your overall reaction to the ITS.

- 1) Overall reaction to the Software:
 - a. Terrible -> Wonderful
 - b. Difficult -> Easy
 - c. Frustrating -> Satisfying
 - d. Inadequate power -> Adequate power
 - e. Dull -> Stimulating
 - f. Rigid -> Flexible
- 2) Screen:
 - a. Reading characters on the screen was . . . hard -> easy
 - b. Highlighting simplifies tasks . . . not at all -> very much
 - c. Organization of information was . . . confusing -> very clear
 - d. Sequence of screens was . . . confusing -> very clear
- 3) Terminology and System Information:
 - a. Use of terms throughout the system was . . . inconsistent -> consistent
 - b. Terminology related to tasks . . . never -> always
 - c. Position of messages on screen . . . inconsistent -> consistent
 - d. Prompts for input . . . confusing -> clear
 - e. Computer informs about its progress . . . never -> always
 - f. Error messages were . . . unhelpful -> helpful
- 4) Learning:

- a. Learning to operate the system . . .difficult -> easy
 - b. Exploring new features by trial and error . . .difficult -> easy
 - c. Remembering names and use of commands . . .difficult -> easy
 - d. Performing tasks is straightforward . . .never -> always
 - e. Help messages on the screen . . .unhelpful -> helpful
 - f. Supplemental reference materials . . .confusing -> clear
- 5) System Capabilities:
- a. System speed . . .too slow -> fast enough
 - b. System reliability . . .unreliable -> reliable
 - c. System tends to be . . .noisy -> quiet
 - d. Correcting your mistakes . . .difficult -> easy
 - e. Designed for all levels of users . . .disagree -> agree
- 6) Please list the most negative aspects of the ITS (identify 3)
- 7) Please list the most positive aspects of the ITS (identify 3)

A.2.3.3 Technology Acceptance Model

Administration Notes

The revised Technology Acceptance Model survey (Davis 1989) is a self-report measure that assesses a users' Perceived Usefulness and Perceived Ease of Use. This measure is typically administered after an introduction or training on a system, and assessed future use based upon initial perceptions. Both portions have been shown to have high internal reliability (Usefulness $\alpha = .98$, Ease of Use $\alpha = .94$), and correlate strongly with usage behaviors. Participants rate the extent to which they agree with each statement using a 7-point Likert scale with anchor point descriptions "Strongly Agree," "Neutral," and "Strongly Disagree." Scores for each factor range from 6 – 42, with higher scores indicating greater perceived usefulness and ease of use.

Sources/references

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS quarterly*, 319-340.

Notes

All items employ a 5-point Likert (response options: strongly disagree, disagree, neutral, agree, strongly agree).

Measures

Perceived Usefulness:

- 1) Using the ITS in COA development would enable me to accomplish tasks more quickly
- 2) Using the ITS would improve my COA development performance.

- 3) Using the ITS in COA development would increase my productivity.
- 4) Using the ITS would enhance my effectiveness in COA development.
- 5) Using the ITS would make it easier to do COA development
- 6) I would find the ITS useful in COA development.

Perceived Ease of Use:

- 7) Learning to operate the ITS would be easy for me.
- 8) I would find it easy to get the ITS to do what I want to do
- 9) My interaction with the ITS was clear and understandable
- 10) I find the ITS flexible to interact with.
- 11) It would be easy for me to become skillful at using the ITS.
- 12) I find the ITS easy to use.

A.2.3.4 System Trustworthiness Scale

Administration Notes

The System Trustworthiness Scale assesses perceived trustworthiness of a system (STS; Muir & Moray, 1996). The STS comprises 9 items, the first six are potential predictors of trust, while the latter three are measures of subjective trust in the system. All items are scored on a scale from 0 (Very Low) to 100 (Very High) in 5-point increments.

Sources/references

Muir, B. M., & Moray, N. (1996). Trust in automation. Part II. Experimental studies of trust and human intervention in a process control simulation. *Ergonomics*, 39(3), 429-460.

Notes

All items employed a response scale from None at all -> Extremely high (responses provided using a slider from 0->100).

Measures

Please use the sliders below to respond to the associated questions.

- 1) To what extent does the system perform its function properly?
- 2) To what extent can the system's behavior be predicted from moment to moment?
- 3) To what extent does the system respond similarly when it encounters similar circumstances at different points in time?
- 4) To what extent can you count on the system to do its job?
- 5) To what extent does the system perform the task it was designed to do?
- 6) Your degree of faith that the system will be able to cope with future events?

- 7) Your degree of trust in the system to respond accurately?
- 8) Your degree of trust in the system display?
- 9) Your overall degree of trust in the system?

A.2.4 Touchpoint Tailored Evaluations

A.2.4.1 SC3 Touchpoint Session 1 ITS Evaluation

Administration Notes

Sources/references

The items in this survey were developed partly in service of the SC3 Touchpoint.

Notes

Items 1-16 employed a 7-point Likert scale (response options: not at all, slightly, somewhat, moderately, mostly, very much, completely). Item 17 employed free response.

Measures

Please respond to the following questions regarding your experience with the Information Technology System (e.g., the combination of the COA-GPT and Simulator Interface). Consider each of the following as “To what extent . . . “

- 1) Did the ITS interface clearly communicate system states and present feedback quickly (e.g., display unit status, highlight key tasks, or ETA for generating COA)?
- 2) Did the ITS follow typical/standard conventions, phrases, terminology, and concepts that are familiar to you?
- 3) Did the ITS follow typical conventions seen in other applications or interfaces you are familiar with (as examples, the Copy function is typically associated with the hotkey [CTRL]+[C]; color conventions often include green= go/on, red=stop/off, yellow=yield/standby)
- 4) Did the ITS check for and warn you about errors (e.g., settings not saved or form fields left empty)?
- 5) Did the ITS provide readily accessible memory aids (such as “tool tips”, reminders, descriptions, FAQ/Help, etc.)?
- 6) Did the ITS promote recognition of information to help you reduce the need to remember to use the system (e.g., Strikethrough often uses the icon “ab” to easily recognize the function)?
- 7) Did the ITS interfaces contain information that is irrelevant, unnecessary or rarely needed?

- 8) Did the ITS display errors messages noticeably in a conventionally recognizable way?
- 9) Did the ITS require additional documentation or explanation to use and, if so, was the documentation provided in context, at the moment you required it?
- 10) Did the sequence of information match your natural planning flow?
- 11) Was the comparison between different COAs easy to interpret and helpful in selecting one?
- 12) Were you able to trace elements of the generated COA back to their sources in the original operations order or input documents?
- 13) Did you verify any of COA-GPT's recommendations independently? What prompted that decision?
- 14) Were Initial CCIRs and EEFI's generated and, if so, were they clear, accurate, and appropriate for the mission-context?
- 15) Did the simulation outputs provide meaningful insight into the likely outcomes of each COA?
- 16) Did everyone interpret the COA-GPT's suggestions the same way?
- 17) If you have any additional feedback or notes from your experience with the ITS please share them below

A.2.4.2 SC3 Touchpoint Session 2 ITS Evaluation

Administration Notes

Sources/references

The items in this survey were developed partly in service of the SC3 Touchpoint.

Notes

Items 1-3 employed multiple choice and free response (response options embedded below for clarity). Items 4a-4q employed a 7-point Likert (response options: not at all, slightly, somewhat, moderately, mostly, very much, completely). Items 5, 6, and 7 employed free response. Items 8a-8c employ 5-point Likert (response options embedded below for clarity). Items 9, 10, and 11 employed free response.

Measures

- 1) How well did COA GPT create a COA that supports [response options: not accurately, slightly accurately, neutral, accurately, very accurately]:
 - a. Problem Statement
 - b. Mission Statement
 - c. Initial Commander's Intent
 - d. Initial CCIRs and EEFI's

- 2) Did you prompt COA GPT to generate new seed COAs multiple times [response options: yes, no]?
 - a. If yes how many times (estimate), and why [response options: free response]?
- 3) Did you manually adjust the tasking or properties of specific units within a given seed COA [response options: yes, no]?
 - a. If yes how many times (estimate), and why [response options: free response]?
- 4) Please respond to the following questions regarding your experience with the Information Technology System (e.g., the combination of the COA-GPT and Simulator Interface). Consider each of the following as “To what extent . . . “
 - a. Was the COA generated from COA-GPT appropriately segmented into phases
 - b. Was the task and purpose for each phase clearly identified?
 - c. Was the information necessary for decision-making both present and clearly understandable?
 - d. Did interacting with ITS clarify any mission planning considerations you hadn’t thought of?
 - e. Were you able to trace elements of the generated COA back to their sources in the original operations order or input documents?
 - f. Were Initial CCIRs and EEFIIs generated and, if so, were they clear, accurate, and appropriate for the mission-context?
 - g. To what extent would you rely on COA GPT for mission plans and COA generation?
 - h. Were Specified and Essential Tasks clearly and accurately identified?
 - i. Were time-based events and phases clearly represented in the interface (e.g., operation start, phase shifts, timing of unit actions)?
 - j. Did the simulation outputs provide meaningful insight into the likely outcomes of each COA?
 - k. Were icons, overlays, and other visual indicators on the map (such as those associated with various units, terrain, Avenues of Approach (AoA), or Defensive positions) clearly and intuitively distinguishable, and doctrinally appropriate?
 - l. Did your team establish any rules or norms for using ITS together (e.g., who enters data, who interprets output)?
 - m. Did everyone interpret the COA-GPT’s suggestions the same way?
 - n. Did team members refer back to ITS outputs to support their arguments or decisions?

- o. Was there a shared understanding of how the team should use ITS during the session?
- p. Did you perceive that COA-GPT and the simulator interface were tools that you and your team used to process and provide information, but did not actively engage in problem solving or decision making as a teammate
- q. Did you perceive that COA-GPT and the simulator interface were contributing as teammates that had responsibility for problem solving and were engaged actively in the team's decision making process
- 5) Did you or your teammates verify any of COA-GPT's recommendations independently? If so, what prompted that decision? [response options: free response]
- 6) Did ITS outputs prompt any disagreements or discussions among team members? If so, what were they about? [response options: free response]
- 7) How did the use of the technologies influence the sequencing and timing of your planning activities compared to when you operated the ITS solo? [response options: free response]
- 8) Please use the scales below to respond to the prompts on the left:
 - a. How effective and useful is the ITS in supporting/conducting MDMP, compared to a traditional approach? [response options: not at all, not very effective, effective, very effective, extremely effective]
 - b. How satisfied are you with the ITSs performance in facilitating MDMP? [not satisfied, slightly not satisfied, neutral, slightly satisfied, satisfied]
 - c. How high is the trust factor of the information provided by the ITS? [response options: very low, slightly low, neutral, slightly high, very high]
- 9) In a future mission, would you prefer to use ITS individually or as a team? Why?
- 10) Would you advise integrating ITS into formal team structures or command roles in future missions? Why or why not?
- 11) Please explain any reasoning for your opinions of the ITS as reflected above

A.2.4.3 SC3 Touchpoint Session 3 ITS Evaluation

Administration Notes

Sources/references

The items in this survey were developed partly in service of the SC3 Touchpoint.

Notes

Items 1a-1k employed a 7-point Likert scale (response options: not at all, slightly, somewhat, moderately, mostly, very much, completely). Items 2, 3, 4, 5, and 7 employed free response. Item 6a-6c employed 5-point Likert scales (response options embedded below for clarity).

Measures

- 1) Consider each of the following as “To what extent . . . “
 - a. Did using COA-GPT to refine elements of a given seed COA proceed as you expected
 - b. Did using COA-GPT to refine elements of a given seed COA produce results that met your needs
 - c. Did using COA-GPT to refine elements of a given seed COA help you achieve a more successful COA
 - d. Did COA-GPT exchanges/requests support your ability to make well-informed timely decisions, relative to instances where you might perform these tasks in a typical, equivalent mission context (using analog tools or present-day technologies, e.g., CPCE)?
 - e. Did you feel that COA-GPT made any assumptions you disagreed with?
 - f. To what extent would you rely on COA GPT for mission plans and COA generation?
 - g. Did you believe that the terrain analysis and unit deployment within COA-GPT outputs (as visualized/displayed in the simulator) to align with mission requirements and constraints
 - h. Were time-based events and phases appropriately managed by COA-GPT and represented in the simulator (e.g., operation start, phase shifts, timing of unit actions)?
 - i. Did the simulation outputs provide meaningful insight into the likely outcomes of each COA?
 - j. Did results from the constructive simulation alter your trust or confidence in specific seed COAs generated by COA-GPT?
 - k. Did you feel that the simulation’s projected results were valid reflections of operational dynamics and/or possible outcomes
- 2) Did you verify any of COA-GPT’s recommendations independently? What prompted that decision?
- 3) Please briefly describe whether you felt confident in the changes to the seed COAs that you made using COA-GPT, and why.
- 4) What impact do you think the full, combined ITS’s outputs, results, outcomes, metrics, etc. had on your planning process compared to traditional COA development? Please consider commenting on the speed

and perceived usefulness of SC3 Sim's feedback compared to traditional methods of estimating COA outcomes.

- 5) What interface or visualization feature do you think contributed most to your feeling of whether you could trust and make use of the seed COAs as visualized in the simulator? (note: that could be a positive or negative impact)
- 6) Please use the scales below to respond to the prompts on the left:
 - a. How effective and useful is the full ITS in supporting/conducting MDMP, compared to a traditional approach? [response options: not at all, not very effective, effective, very effective, extremely effective]
 - b. How satisfied are you with the ITSs performance in facilitating MDMP? [response options: not satisfied, slightly not satisfied, neutral, slightly satisfied, satisfied]
 - c. How high is the trust factor of the information provided by the ITS? [response options: very low, slightly low, neutral, slightly high, very high]
- 7) Finally, please provide any other feedback you would like to see regarding the ITS, the touchpoint sessions, the general direction of the research effort, or anything we have not touched on this week?

A.2.4.4 System Elements Evaluation

Administration Notes

Sources/references

The items in this survey were developed partly in service of the SC3 Touchpoint.

Notes

All items employed free response.

Measures

- 1) What Integrated Technology System (ITS) features were useful and how? Please try to comment on the specific elements listed below:
 - a. What specifically about the COA-GPT component?
 - b. What specifically about the constructive simulation component?
 - c. What specifically about the user interface component?
- 2) What ITS features were problematic and how? Please try to comment on the elements listed below:
 - a. What specifically about the COA-GPT component?
 - b. What specifically about the constructive simulation component?
 - c. What specifically about the user interface component?

- 3) Which element of the ITS is most lacking, specifically in terms of hindering your ability to conduct planning in a sufficiently realistic manner (e.g., gaps in representations, oversimplifications, inaccuracies, etc.)? Please consider what you would say are the most critical of these lacking areas to address and in what way would you like them addressed?

A.2.4.5 MCBL BOLT Questionnaire (Modified)

Administration Notes

Sources/references

The items in this survey were developed partly in service of the SC3 Touchpoint.

Notes

Items 1-11 employ a sliding response scale from 0->100 labelled at five intervals (labels: no improvement, slight improvement, neutral, very great improvement, extreme improvement). Item 12 employed free response.

Measures

Please indicate to what extent you feel that the following design features would improve the ITS's ability to support your iteration towards an ideal COA and/or to work on that iterative process as part of a team.

- 1) The ability to copy-paste properties and/or tasks from one unit (or set of units) to another
- 2) The ability to draw on the mission map to designate important geometries such as regions that are the "domain" of one or another unit or perhaps represent avenues of approach, etc.
- 3) The ability to manage units at different echelons and to designate/create groups of units that may be given the same sets of tasks all at once rather than having to set those tasks individually for each unit
- 4) The ability to use COA-GPT to engage in additional, natural language iterations that allow you to modify specific components of a given seed COA (e.g., to use natural language to alter the COA for a specific unit such as C 1-1 or for a given phase of the mission rather than modifying their plan manually through the interface)
- 5) The ability to manage COA phases distinctly such that one could modify the plans for units or sets of units up to a certain time or phase line, and then proceed to adjusting plans for the next phase
- 6) The ability to run the simulation of the COA in phases or up to certain time points to allow you to inspect when/where events occur and issues arise

- 7) The ability to view unit anchored menus within the map view that allow you to specify unit properties and tasks within that view rather than using the taskbar and drop-down menus (that is, the UI tools on the left)
- 8) The ability to toggle granularity of resource costs in the simulator outputs to allow you to drill down into COA costs
- 9) The ability to set unit positions for MoveTo tasks by dragging and dropping units and designating the time (or mission phase) by which they should reach the given destination
- 10) The ability to toggle back and forth between the full overview of the operational area and a zoomed in view such that one may be zoomed into and altering the tasking for a given unit, toggle to zoom out to the full view to review positions, and then toggle again to return to the unit one was inspecting
- 11) The ability to war game against another human who is tasked with developing a COA for the enemy forces (note: this capability would perhaps depend on being able to run the simulation and re-plan in phases or at critical decision points)
- 12) Finally, please use this section to elaborate or to recommend other capabilities/features that you think would improve the COA development and testing process. Alternatively, please comment on any additional feedback or notes from your experience with the ITS

...

List of Symbols, Abbreviations, and Acronyms

1-1CAV	1st Squadron, 1st Cavalry Regiment
1-6IN	1st Battalion, 6th Infantry Regiment
1/2ID	1st Brigade Combat Team, 2nd Infantry Division
2/1AD	2nd Brigade Combat Team, 1st Armored Division
25ID	25th Infantry Division
3D	Three-Dimensional
5-20IN	5th Battalion, 20th Infantry Regiment
715 BCG	715th Brigade Combat Group
AAR	After-Action Review
AI	Artificial Intelligence
AO	Area of Operation
AOI	Area of Interest
ARL	Army Research Laboratory
C 1-1	C Company, 1st Squadron, 1st Cavalry Regiment
C2	Command and Control
COA	Course of Action
COL	Colonel
CONOPS	Concept of Operations
DEVCOM	U.S. Army Combat Capabilities Development Command
FOE	Future Operational Environment
FY	Fiscal Year
GPT	Generative Pretrained Transformer
GUI	Graphical User Interface
HCxS	Humans in Complex Systems
HE	Higher Echelon
ITS	Integrated Technology System

LLM	Large Language Model
LOS	Line of Sight
LSCO	Large-Scale Combat Operation
MCBL	U.S. Army Mission Command Battle Laboratory
MDMP	Military Decision-Making Process
ML	Machine Learning
NASA-TLX	National Aeronautics and Space Administration Task Load Index
OBJ	Objective
OPORD	Operations Order
PAA	Position Area for Artillery
SC3	Scalable Cross-Echelon Command and Control
SME	Subject Matter Expert
STP	Soldier Touchpoint
UCF	University of Central Florida
UI	User Interface

1 DEFENSE TECHNICAL
(PDF) INFORMATION CTR
DTIC OCA

1 DEVCOM ARL
(PDF) FCDD RLB CB
TECH LIB