ARL-TR-10391 • JULY 2026

A Method of Assessing Basic Research in Autonomy

by Craig Lennon

NOTICES

Disclaimers

ARL-TR-10391 ● July 2026

A Method of Assessing Basic Research in Autonomy

Craig Lennon
DEVCOM Army Research Laboratory

REPORT DOCUMENTATION PAGE

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED START DATE	END DATE
July 2026	Technical Report	1 March 2026	30 April 2026

4. TITLE AND SUBTITLE
A Method of Assessing Basic Research in Autonomy

5a. CONTRACT NUMBER	5b. GRANT NUMBER	5c. PROGRAM ELEMENT NUMBER
5d. PROJECT NUMBER	**5e. TASK NUMBER**	**5f. WORK UNIT NUMBER**

6. AUTHOR(S)
Craig Lennon

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)	8. PERFORMING ORGANIZATION REPORT NUMBER
DEVCOM Army Research Laboratory ATTN: FCDD-RLA-JC Aberdeen Proving Ground, MD 21005	ARL-TR-10391

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
ORCID: Craig Lennon, 0000-0002-2229-2919

14. ABSTRACT
This report builds on the discussion in *Framework for Assessing Autonomy Research* (ARL-TR-10064, Mar 2025) and proposes a method by which basic research in autonomy can be assessed for its potential application to support U.S. Army missions. The method was demonstrated in experiments conducted by the Distributed and Collaborative Intelligent Systems and Technology Collaborative Research Alliance in April 2026, and that set of experiments is described here to illustrate the method.

15. SUBJECT TERMS
Military Information Sciences, autonomy, artificial intelligence, AI, robotics, experimentation

16. SECURITY CLASSIFICATION OF: a. REPORT	b. ABSTRACT	C. THIS PAGE	17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UU	38

19a. NAME OF RESPONSIBLE PERSON	19b. PHONE NUMBER *(Include area code)*
Craig Lennon	(520) 691-6117

STANDARD FORM 298 (REV. 5/2020)
Prescribed by ANSI Std. Z39.18

Contents

List of Figures

List of Tables

1. Introduction and MIT Research

Evaluating basic research in autonomy is challenging for many reasons, with the following amongst them:

- Autonomy is defined as a system-level property (Hicks 2023), so we want to evaluate the contribution of the research to system performance—but it is generally not practical to implement basic research on Army systems.
- The performance of autonomy depends on the operational environment in which it is deployed, so we want to evaluate autonomy research across varied environments to ensure it is robust—but the scientific method requires control of environmental properties relevant to the research.
- How an autonomous system acts depends on the sequence of previous actions taken by the system, so it should be evaluated in the context of a mission—but the contribution of a single research project implemented as a component of the system is difficult to evaluate when system performance can be affected but any one of a sequence of previous actions in the mission.

The framework in Lennon (2025b) proposed a series of steps for evaluating autonomy research, addressing at a high level how the first two challenges might be overcome. This report shows how that framework can be applied to basic research, and how the third challenge might be overcome. The application is to an experiment conducted in April 2026 by the Distributed and Collaborative Intelligent Systems and Technology (DCIST) Collaborative Research Alliance (CRA) that has produced basic research that U.S. Army Combat Capabilities Development Command (DEVCOM) Army Research Laboratory (ARL) is assessing for its potential application to Army operations. Those April 2026 experiments built on research documented in Lennon (2025a). This report describes the evaluation of research by the Massachusetts Institute of Technology (MIT) into creating a scene graph and using it as the basis for translating natural language commands to language usable be a robotic planner and grounding the resulting plan to the robot's environment. The MIT research was described in Section 3 of Lennon (2025a); thus, this report provides only a brief description of that research at the end of this Section and instead focuses on the methods used to evaluate the research.

The April 2026 DCIST CRA experiment was based on a scenario general enough to be the context for evaluating several different DCIST CRA research projects. Within this scenario, research teams were given a series of missions, each of which required robots to accomplish one or more tasks, such as finding a target object or blocking a road. The initial conditions of the task varied in order to evaluate the robustness of the research, and performance was evaluated by grading how the actions of the robot contributed towards subtasks into which the task was decomposed. Researchers spent several days at one site, adapting research previously tested at their home university to the new environment at Camp Buckner (West Point, New York). The Camp Buckner environment in April 2026 differed not just from their own institutions, but also from their previous experience at Camp Buckner in Lennon (2025a), which took place in November 2024. After adapting to one site at Camp Bucker, the researchers conducted missions at a new site at Camp Bucker to evaluate the robustness of their adaptations.

Rather than repeating the overview of the relevant MIT research project presented in Lennon (2025a, Section 3), the following focuses on properties of the research relevant to this experiment. The scene graph was a data structure that contained hierarchical information relating relevant semantic objects and properties of their environment. For example, an object like a cone could be detected during the mapping process and mapped to a place, which is a point in the world. Regions are collections of places, with examples being road intersections and parking lots. Other objects besides cones were mapped, and the ones relevant to this report are buildings and cars. At the time of the experiment, the scene graph was not updated online during runs, which means the robots could not explore new areas or even recognize new objects in previously explored regions during the run.

A large language model (LLM) took commands and translated them into goals described in planning domain definition language (PDDL), then grounded the objects referenced in the goals to objects in the scene graph. Because PDDL is not temporal, goals cannot express sequential events such as "go to point A and then go to point B," nor is time to complete the mission considered during the planning process. At the time of the experiment, the commands needed to be presented directly to the robots. An example follows, in which "Hamilton" and "Euclid" are the names of robots.

"Tell Euclid to use a cone to block the intersection near the cones. Tell Hamilton to go to the cars in the parking lot closest to the barracks."

Intersections and parking lots are regions within the scene graph, and the LLM would reason through which of these regions were closest to the cones and barracks, respectively. After identifying the regions, the LLM would select a cone to be placed at a point within the intersection by Euclid and a car within the parking lot to be visited by Hamilton.

Finally, the only application of AI off-platform was the use of the cloud-based language model to translate natural language to PDDL goals and ground them to objects or regions within the scene graph. While other research teams used an off-platform visual-language model to describe activity or objects (see Lennon [2025a], Section 4), the MIT team did not and therefore did not describe activity around target objects.

Section 2 describes the scenario used for this experiment and presents an example mission along with the tasks required of the robot during the mission. Section 3 describes how tasks relate to subtasks, how observable actions of robots were mapped to subtasks for evaluating the MIT research, and presents the results of the evaluation. Section 4 draws conclusions about the research and the method of experimentation by which the research was conducted.

2. Scenario

The scenario for an experiment is the fictional but Army-relevant situation that implicitly defines what types of missions and operational environments, and therefore what types of tasks and conditions, are within the realm of possibility for the experiment. The scenario is agreed upon in advance of the experiment and should be rich enough so that all researchers in the experiment can propose robot behaviors enabled by their research that are relevant to the scenario. We referred to the scenario for the April 2026 experiments as "Find Mr. X," as it centered around finding a target person or vehicle within a defined area of Camp Buckner. A scenario is analogous to the situation of an operations order (OPORD), which lays out the area of operations, friendly and enemy forces, and other relevant information that might be useful for a variety of possible missions. Appendix A contains the OPORD provided to teams for this mission, which had one situation common across research teams and several missions of progressing difficulty for each team.

While each research team operates withing the same scenario, they have different and changing missions. A mission is the context within which a research team conducts a run of the experiment. Two research teams may have the same mission but execute it differently. An example mission follows, with items in brackets indicating information that is filled in at the time the mission is given.

"Starting at [start location] at [start time], your team will conduct a reconnaissance of Area Alpha in order to find Mr. X's vehicle and send back a report describing the vehicle and activity around it. Complete the search no later than [stop time]. The car was last seen at [location]."

Though the same mission could be given to any research team, the methods and capabilities of the teams differ, which may result in the mission being refined or changed in scope before instructions are provided to robots. For example, in the case of the MIT research described in Section 1, the start and end times were not transmitted because PDDL does not consider time. No report describing activity around the car was expected because the robot did not have the capability for that, and the location of the car had to be a location at which a car had previously been detected since the robot did not update maps online. Thus, the task passed to the robot for this mission was simply going to the car, with instruction to the LLM as follows.

“Mr. X’s car is located between the hospital and the government building. Tell Hamilton to go to his car.”

Other research teams might take the same mission and present different tasks to their robots depending on those robots’ capabilities. The mission is an Army-relevant team activity that provides the opportunity for robots to perform tasks demonstrating the applicability of their research. But the mission is interpreted and passed as tasks to the robots in a form consistent with their capabilities, just as a Squad Leader would present role-relevant instructions to Soldiers instead of just passing along the platoon OPORD.

The scenario provides a general context for an experiment, and missions within the scenario present research teams with the opportunity to assign their robots to tasks appropriate for the capability of the team’s robots. These tasks are the basis for evaluating the research, which we address in Section 3.

3. Tasks and Evaluation

The task to which a robot is assigned will be accomplished by a sequence of actions that, as noted in the introduction, make it challenging to evaluate the contribution of research late in the sequence of actions. Thus, there is an advantage to breaking the task into a sequence of subtasks; if the robot fails at some action in the sequence, it can be assisted through that subtask and graded on the next subtask in the sequence as if the previous one had been correctly done. How to decompose the task into subtasks depends on how the mission is expected to run, the environment, the research to be assessed, and how that research is implemented on the demonstration platform.

Different tasks may share common subtasks. Table 1 shows the decomposition of two tasks into subtasks. Subtasks are listed in the left column, with the tasks (blocking an intersection and identifying a target) listed across the top row. The cells in the column below each task are noted with an “X” if the subtask was part of the task.

Table 1. Example task decomposition into subtasks.

Subtask (down)	Block intersection	Identify target
Map region	X	X
Planning	X	X
Search area	X	X
Inspect object	. . .	X
Block road	X	. . .

The subtask is the level at which a research team may be evaluated, but the evaluation itself is based on observable actions of the robot, with success criteria defined by the research-related actions performed during this subtask. We are trying to evaluate the research using observations of actions taken by the robots, with the subtask constituting a step of the task that can be graded separate from other subtasks. We are not trying to evaluate the subtask on the basis of whether it was accomplished as would be needed in a tactical environment. For example, if a research team was trying to demonstrate the ability to hide from the enemy, maintaining concealment would be an action related to one or more subtasks, but they would not be evaluated on maintaining concealment if they are not engaged in such research. For the MIT team, the key research was using a scene graph and LLM to ground natural language to objects and locations, so actions related to each subtask were as follows:

- Mapping the region by creating a scene graph offline.
- Planning by creating a correct PDDL goal and correctly grounding it.
- Searching an area by navigating to a position corresponding to a previously mapped object.
- Inspecting an object by going to the correct object and relaying that it had reached the correct place.
- Blocking a road by retrieving a cone and placing it at an intersection.

Table 2 shows the mapping of actions to subtasks. We now turn to evaluation of the actions and the results of that evaluation.

Table 2. Mapping of actions to subtasks.

Subtask	Action for MIT team
Map region	Create scene graph of the area
Planning	Ground to the correct region
	Ground to the correct object type
	Express correct goal in PDDL
Search area	Navigated to the correct region
	Navigate to the correct object
Inspect object	Move to target object and report as inspected
Block road	Pick up cone and place it in the intersection

For most actions in Table 2, it is clear how you would evaluate the action, excepting possibly the creation of the scene graph. The scene graph was evaluation based on correct classification of terrain traversability over the route relevant to the task, and whether the mission-relevant objects were mapped. Before the experiment, general evaluation criteria were created for each subtask, which the experiment designers imagined would cover any approach to the mission. These were refined based on the capabilities of the research team at the time of the experiment. The general evaluation criteria are included in Appendix B. To illustrate what was changed for the MIT evaluation, the general evaluation criteria for searching an area are as follows:

- Robots moving in the area correctly identify relevant objects.
- Structures mentioned in the mission are identified via symbol.
- Cones used for blocking roads identified (if mission relevant).
- Vehicles of relevant color identified.
- People identified (if searching for a person).
- Vehicles matching Mr. X's description are inspected, with the robot moving around the vehicle to look for the identifying symbol.
- If all objects in the area are identified as neither the target object nor a mission-relevant object, the search of the area stops but the mission continues.

The MIT scene graph mapped vehicles but did not record their color and did not recognize signs or symbols on buildings, nor people and their manner of dress. Thus, searching an area was modified to the following:

- Identified relevant objects.
- Identified relevant structures.
- Identified cones.
- Identified vehicles based on their described position.
- Navigated to the correct object, meaning it navigated to the correct area and to the correct object within that area.

The first four bullets were graded as part of the mapping evaluation, meaning that the only actions evaluated for search were navigating to the correct area and correct object within the area it went. The general evaluation criteria relate to important aspects of the subtask, but since the research was being evaluated, we graded research-related actions and mapped them to subtasks based on what best facilitated evaluation using the general evaluation criteria as a guide. The right column of Table 2 is what was graded for the MIT evaluation.

As the research team completed tasks, they were provided with more difficult tasks wherein task difficulty had been established before the experiment, based on expectations of what would present a challenge for the teams. Task difficulty for the MIT evaluation progressed as follows:

1) A single robot sent to an object, for example, “Mr. X’s car is located between the hospital and the government building. Tell Hamilton to go to his car.”
2) Robots are sent to different locations, for example, “Mr. X’s vehicle is a car or a truck. It is in one of two places between the barracks and the hospital or in the parking lot near the government building. Tell Euclid to go to the vehicles in the parking lot. Tell Hamilton to go to the vehicle between the buildings.”
3) Robots are sent to multiple objects, for example, “Mr. X’s car is at one of two locations in the parking lot near the barracks or between the hospital and the government building. Tell Euclid to go to the cars in the parking lot near the barracks. Tell Hamilton to go to the car between the hospital and the government building.” Euclid was expected to visit more than one car.
4) Block intersection, for example, “Tell Euclid to use a cone to block the intersection nearest the government building. Have Hamilton head to the cars in the government building’s parking lot.”
5) Robots block distinct intersections, for example, “Tell Hamilton to use a cone to block the intersection near the government building. Tell Euclid to use the other cone to block the intersection near the barracks.”
6) Robots redirected by fragmentary order, in which case there are two orders, such as “Mr. X’s vehicle is a car or truck. Tell Hamilton to go to the cars in the parking lot by the government center. Tell Euclid to go to the truck between the barracks and the hospital.” Then a second order is issued, such as “We found Mr. X in the parking lot. Tell Euclid to help Hamilton.”

There were two areas in which teams conducted missions, both shown in Figure 1.

Figure 1. Area Alpha outlined in green and Area Bravo outlined in gold.

The MIT team started in Area Bravo and stayed there for 3 days before moving to Area Alpha to spend an additional 2 days. Task no. 2 of the six previously listed tasks took place in Area Alpha. Figure 2 shows a closer view of the area within Alpha traversed during the execution of this task. The key buildings are labeled with blue tags, and the cars (two in the parking lot and one between the barracks and hospital) are labeled as objects 100, 101, and 106. The area shown Figure 2 is also where task no. 5 was executed, with robots retrieving cones from the area near the government building marked with an orange tag and placing them at the two intersections marked with purple tags. Figure 3 shows one of the two similarly configured robots used by MIT for the evaluation.

Figure 2. Example of key locations for two tasks.

Figure 3. Example robot used in the experiment.

The scoring of each of the actions listed in Table 2 was on a scale from 1 to 3 and is shown in Figure 4 according to the following system:

- Fully autonomous performance of an action is dark green and scored as 3.
- Minor assistance from the operator is light green and graded as 2.
- Substantial assistance is yellow and graded as 1.
- When the action was not attempted, either because it was not required for the task or because the robot did not perform it for other reasons, the cell is left blank (white). A blank cell does not count as success for failure, because not all actions were required for a given mission. For example, most missions did not require blocking the road.

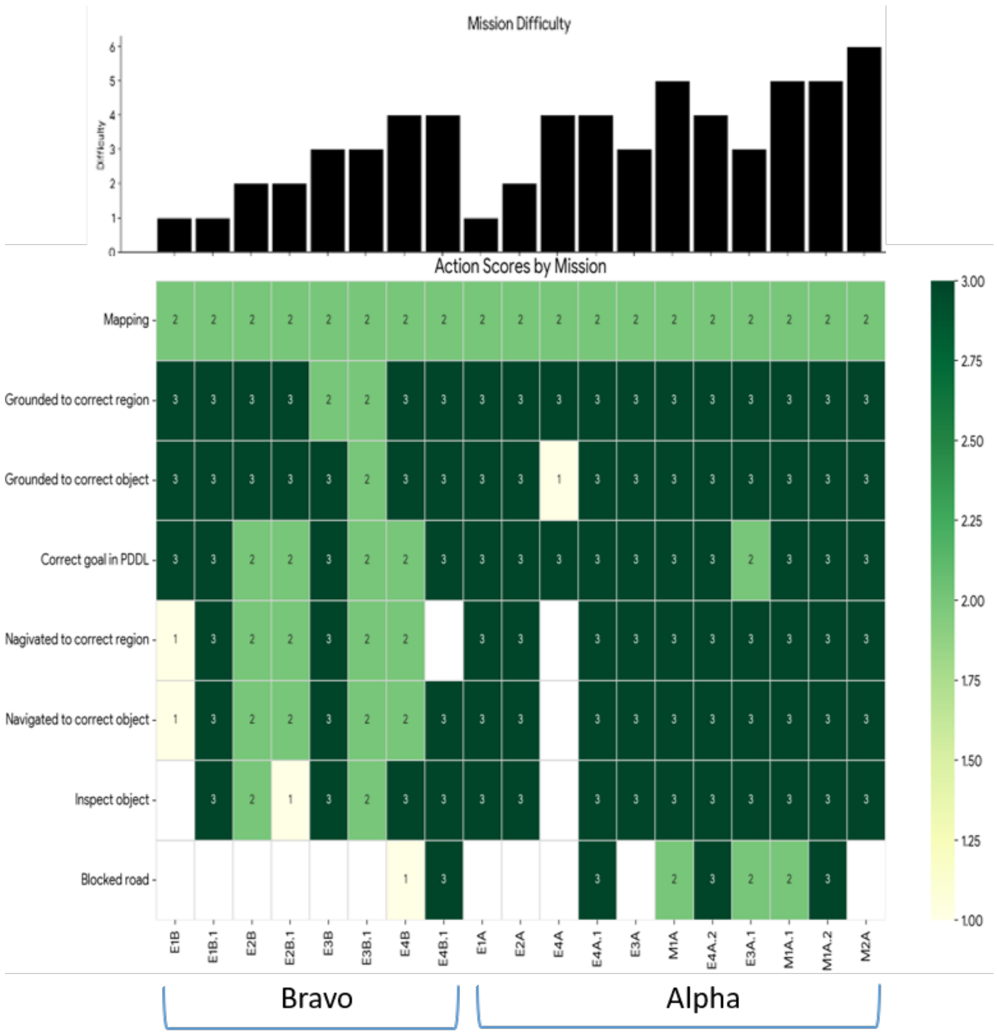


Figure 4. Action score for each mission by level of difficulty from 1 as worst to 3 as best.

Grading by action rather than success or failure for the task allows the robot to perform poorly on some actions but receive operator assistance to finish the subtask and then demonstrate its ability to achieve success on actions corresponding to subsequent subtasks.

Each mission was given a unique identifier which is shown on the horizontal axis of Figure 4's heatmap, for example "E1B" or "E1B.1." The system behind the identifier is not relevant for interpreting results, so they may be considered as names of runs. At the top of Figure 4 is the difficulty for each run, and the actions for each run fill the vertical axis of the heatmap, with each cell showing the score for the action by run. Below the heatmap, dividers show that 8 runs took place in the first 3 days at Bravo, with 11 more over the subsequent 2 days at Area Alpha. Figure 5 shows how researchers classified the reason for failure or intervention. Planning and navigation errors over the first four runs were attributed to a combination of difficulty with the traversability map, midrange planner, and localization. One camera in the hand of a robot was broken, which prevented grasping cones, but was eventually replaced. The LLM required additional context to ground successfully on the runs marked in green, and the "other" category is attributable to communications related failures.

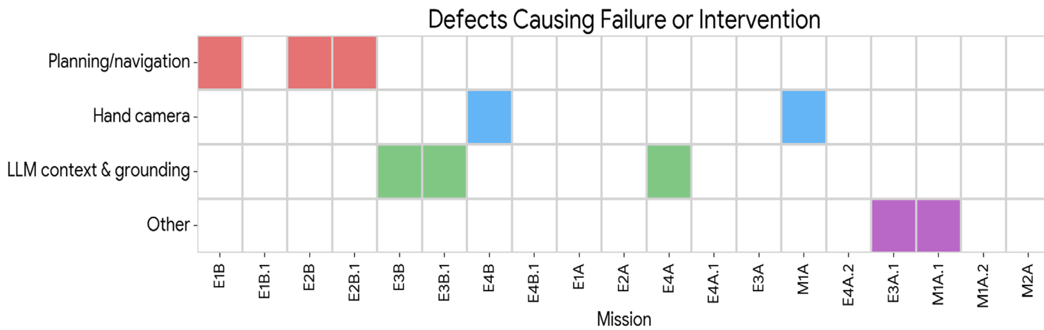


Figure 5. Classification of defects causing failure or intervention.

The key observations from these results are as follows:

- As the scene graph was not updated during runs, it needs regular human editing when new objects are introduced into the environment.
- After three days of fixing bugs and adapting the system in Area Bravo, the runs in Area Alpha demonstrated fast progress, indicating that the adaptation was robust to changing the area of operations.
- There was difficulty navigating during the first four runs, which was attributed to a combination of mid-level planning, traversability, and localization. This was compensated for by editing the scene graph to indicate which areas were not traversable, after which navigation improved. However, the process of editing the scene graph for traversability also limited the distance over which the robot could travel, as longer missions required more editing, so while missions of 300–400 m were included in the original experiment plan, the maximum mission length executed in April 2026 was approximately 150 m.
- Choice of planning language limits mission capability. For example, the planner could not plan sequential tasks since PDDL is nontemporal.
- Prompt engineering is still a critical factor in the process, as indicated by the LLM-related defects in Figure 5.

Section 3 draws conclusions related to the method of experimentation and research that was the subject of the experiment.

3. Conclusions

With respect to the method of experimentation, the general approach was sufficient for the April 2026 DCIST CRA experiment. By decomposing tasks into subtasks, changing the areas of operations halfway through the experiments, and introducing a variety of missions with different initial conditions and progressively increasing difficulty, we mitigated the challenges to evaluating autonomy research described in Section 1. We will continue refining this method in the DCIST CRA Capstone experiments. The scenario, evaluation criteria, and other preparation may have been too extensive for what was done at the experiment. The full OPORD in Appendix A was not needed by any of the research teams, which only used tasks derived from the mission statement. The general evaluation criteria in Appendix B were useful but needed to be refined into the observation of particular actions. Each planned run has setup instructions and an evaluation sheet (see Appendix C for an example) but these had to be modified on-site to deal with changing conditions and were not used during the experiment. It might have been better to lay out general setup guidance rather than plan out each run.

Regarding the research, some failures during runs are fixable by an improved traversability map. As traversability was not a focus of this research project, this issue—like those related to broken cameras and communications—does not provide insight into the benefits and limitations of the basic research. There is insight to be gained from the LLM context failures. The scene graph data structure was created for urban areas and therefore has a hierarchy of places, rooms, buildings, etc. When moving to a new area, the relation of hierarchy layers needs to be transferred, and this is likely a decision that a human will need to make and inform the LLM of it. For example, in one run, the LLM needed to be provided context that "the parking lot is a room." The parking lot was not a room, but the operator was reasoning through the LLM errors and modifying context so as to cause the LLM to identify the appropriate tier of the data structure hierarchy with its equivalent in the mission environment. When applying LLM as part of autonomy to facilitate tasks, we should anticipate that a human supervisor may need to "explain" to the model how the current operational environment relates to the model's current context or prior learning and thus would need to understand the model's reasoning and current context well enough to do so.

4. References

Lennon C. Distributed and collaborative intelligent systems and technology collaborative research alliance fall 2024 integration testing results. DEVCOM Army Research Laboratory (US); 2025a Sep. Report No.: ARL-TN-1274.

Lennon C. Framework for assessing autonomy research. DEVCOM Army Research Laboratory (US); 2025b Mar. Report No.: ARL-TR-10064.

Office of the Under Secretary of Defense for Policy. DoD Directive 3000.09 autonomy in weapon systems. Department of Defense (US); 2023 Jan. https://www.esd.whs.mil/portals/54/documents/dd/issuances/dodd/300009p.pdf

Appendix A. Experiment Operations Order*

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

Situation

Your team is part of the First Robotic Task Force, currently in Camp Buckner, NY, which is designated by the bounded region on your map. The town of Camp Buckner, NY is politically divided between two competing factions, the Mule party and the Knight party. Tensions between the two factions are high due to the current Mayor, who represents the Knight party, postponing recent elections, claiming that they could not be held on schedule due to a cyber-attack which compromised the town government's computer network. The leader of the Knight party, Mr. X, claims that the Mayor is no longer legitimately in power, while the Mayor claims that elections will occur as soon as the network has been fixed.

The Mule party maintains friendly relations with your institution, while the Knight party is neutral towards your institution. The Knight party does maintain friendly relations with an opponent of your institution, which we will simply refer to here as your opponent. Your opponent has its own robotic team, which may be in the area. The type and number of robots they have on their team is unknown.

The mayor of Camp Bucker has requested assistance from your institution, and your institution has agreed to assign your team to assist the mayor in keeping order in the town. Your team operates under the following rules of engagement.

1) Your team members may not directly harm any person or animal of significance to a person (e.g. pets, livestock) during operations.
2) Your team may observe, record, and transmit or save data from any point within the town.
3) Your team may block roads when directed, by placing a robot, cone, or other obstacle in the road.
4) Robots in the town which are not part of your team may belong to your opponent. You are not required to take any action toward them, but should report them when you observe them. Do not initiate any hostile action towards your opponent's robots unless directly told to do so in an order.
5) You should avoid damaging property within the town unless doing so is essential to completing your mission. Ask permission before taking any such action which damages property unless you need to take the action to adhere to rule 1.

The Buckner Militia and two of the town's important strategic facilities, a communication center and a warehouse, are controlled by the Knight party. The Buckner Militia wears camouflage uniforms around town, and you should report persons wearing camouflage uniforms as potential members of the Bucker Militia. The warehouse is run by the Buckner Warehouse Union (BWU), whose members wear red vests. If you see a person wearing a red vest, you should report them as a potential member of BWU. Persons working in the communications center do not wear camouflage and do not wear vests. Mr. X is distinctive in that he always wears a green vest, and you should report anyone in a green vest as Mr. X.

While the Buckner Militia and BWU are loyal to the Knight party, the communications center has some members who are willing to assist the Mule party by providing information, either by electronic means or by physical signal, to your team.

Figure A-1. Version Alpha of Buckner.

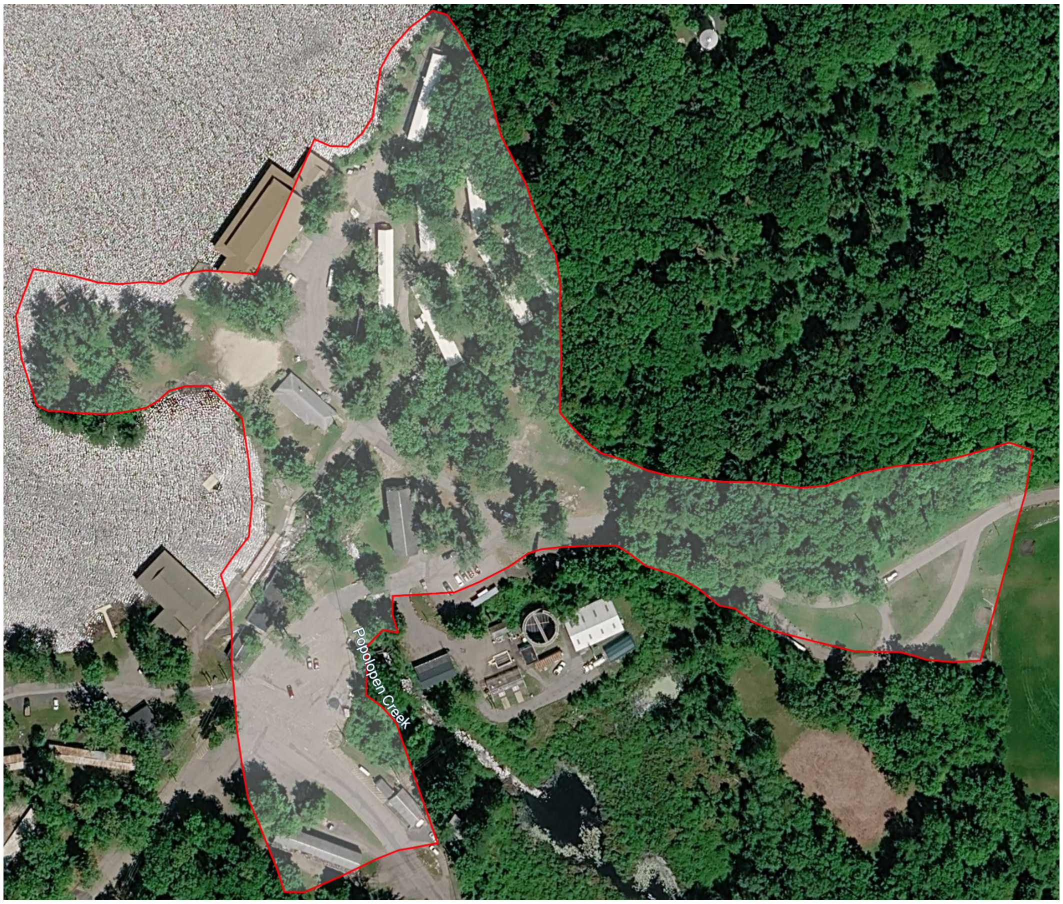


Figure A-2. Version Bravo of Buckner.

Mr. X may try to take control of the town by one of the following courses of action. The first course of action Mr. X might pursue would be to have the Buckner Milita physically take control of the town government, detaining its current employees and replacing them with those loyal to the Knight party. With the town administrative and legal structure under his control, his party could hold an election and bring the town under formal political control. This course of action would be initiated by a movement of the Buckner Militia from their barracks to the government buildings. In this course of action, it is likely that Mr. X would shut down all communications in the town until the Militia has completed its task. This means Mr. X would need to physically go to the Barracks to coordinate the action. Your team's source in the communications center have agreed to set a traffic cone out in front of the communications center if Mr. X is headed to the Barracks area, or to send a message with that information depending on which they can do.

The second course of action would be to use members of the BWU to threaten, intimidate, and harass the administrative government personnel until they agree to hold elections under the supervision of the BWU. Under cover of protesting government policies, the BWU could threaten violence against members of the government loyal to the Mule party and Mr. X could keep local law enforcement from intervening with the threat of deploying the Buckner Militia. This could involve direct armed confrontation of local law enforcement by the Militia, but could also be as simple as the Militia blocking off roads and other avenues of approach that would allow law enforcement to stop the BWU. For this method to work, the Knight party would likely need to ship weapons into the town in order to arm the BWU. Roadblocks set up by the Bucker Militia or weapons shipped to the warehouse are indicators of this course of action.

In both courses of action, control over the town communications facilities is crucial for the Knight party for either cutting off communication or making sure their own message gets out.

In addition to controlling the political leadership and civilian government apparatus, the Mule party also controls the Buckner Maintenance Union (BMU) which handles town maintenance, construction, and sanitation. The BMU is distinctive for its orange vests, and in the event of factional conflict would likely try either to take control of the communication center or block avenues of approach to and from the barracks to stop the Buckner Militia from acting.

Mission

The sequence of missions is at the end of the order.

Execution

Your team will assemble at your start position and begin the search of your area at the time you are directed to start. Other elements of your task force are searching outside your area, so you should stay within the bounds of the area and report if you need to leave that area.

Mr. X's car is [color or other description], with is distinguishable by a [symbol to be provided] on the car.

Mr. X himself is identifiable because he will be wearing a green vest.

If you encounter roadblocks, do not pass them by going off to the side of the road around them. Report them and find an alternate route on another road if possible, or request assistance if an alternate route is not available.

Communication

Report leaving your start location to begin the search, and report when your mission is complete. Reports on those actions, and other actions you take such as blocking roads may be in any format, but should include the time at which the action was taken. Some reports will be requested in SALUTE format, which is size, activity, location, unit or organization, time, and equipment.

Size: The number of personnel, vehicles, or assets involved. This could also include the estimated size of a group if you cannot count all members.

Activity: Describe what the observed entity is doing, including their actions, behavior, and direction of movement. Describe anything related to the situation presented in the OPORD.

Location: Identify where the observation took place. By GPS, map location, or other description. If it is near a structure or landmark then report that as well.

Unit or Organization: Identify the unit or group observed, if possible, or describe their uniforms, clothing, or any distinguishing insignia.

Time: Report the date and time when the activity was observed.

Equipment: Describe any equipment, weaponry, tools, or vehicles associated with the observed entity, including make, model, color, and any unique markings.

Use this SALUTE format when directed in the mission or when it is appropriate to your circumstances. Make sure to report the following if you see them, in SALUTE format.

Report persons wearing camouflage as members of the Buckner Militia

Report persons wearing red vests as members of the BWU

Report persons wearing orange vests as members of the BMU

Report a person wearing a green vest as Mr. X.

Report roadblocks if observed.

Report a cone placed in front of the communications center if observed.

Report Mr. X’s car if you observe it.

You are not required to make recommendations or predictions, but if you have recommendations or predictions you think might be helpful based on the context provided in the OPORD and your observations, provide them in the following format:

Recommendation: Briefly state the action you recommend taking to accomplish the mission or other objective and your reasoning for recommending it.

Predictions: Briefly state what you predict is going to happen or briefly state your prediction about something you have not yet observed. State the reason for your predictions

Section: Sustainment

If you are unable to continue, or need assistance, report this immediately.

Annex:

Images of relevant symbols, signs, vests provided as needed.

The following is a list of the missions proposed to be executed prior to the experiment, and tasks executed during the experiment if format of alternating mission and task.

Mission E1: Starting at [start] at [time start], your team will conduct a reconnaissance of Area [Alpha or Bravo] in order to find Mr. X's vehicle and send back a SALUTE report describing the vehicle and activity around it. Complete the search no later than [time2]. The car was last seen at [actual location]

Task: Mr. X's car is located between the hospital and the government building. Tell Hamilton to go to his car.

Mission E2: Starting at [start] at [time start], your team will conduct a reconnaissance of Area [Alpha or Bravo] in order to find Mr. X's vehicle and send back a SALUTE report describing the vehicle and activity around it. Complete the search no later than [time2]. The car may be at one of the following locations [initial position, v1, or v2].

Task: Mr. X's car is at one of two locations between the hospital and the government building or near the communication center. Tell Euclid to go to the car between the hospital and the government building. Tell Hamilton to go to the car near the communication center.

Mission E3: Starting at [start] at [time start], your team will conduct a reconnaissance of Area Alpha in order to find Mr. X's vehicle and send back a SALUTE report describing the vehicle and activity around it. Complete the search no later than [time2]. The car may be at one of the following locations [v1 or v2]. If you see a traffic cone in front of the Hospital, search near the government building for Mr. X car.

Note: The above was planned before we understood that there was no real-time activity detection to report objects in front of buildings or around cars.

Task: Mr. X's car is at one of two locations in the parking lot near the barracks or between the hospital and the government building. Tell Euclid to go to the cars in the parking lot near the barracks. Tell Hamilton to go to the car between the hospital and the government building.

Mission E4: Starting at [start] at [time start], your team will conduct a reconnaissance of Area [Alpha or Bravo] in order to find Mr. X's vehicle and Mr. X. Upon finding Mr. X's vehicle send back a SALUTE report describing the vehicle and activity around it. Upon finding Mr. X, send back a SALUTE report describing the vehicle and activity around him. Complete the search no later than [time2]. The car and Mr. X may be at one of positions [initial position, v2].

E4 Fragmentary order.

We have received a report that Mr. X is near the hospital and Mr. X car may be near [v1]. Search these locations and report as in the original mission.

Starting at [start] at [time start], first block the road between the Government building and the barracks. Report when done.

Note: Instead of issuing the fragmentary order, we chose to test whether we could correctly place the cone in the main order.

Task: Tell Hamilton to use a cone to block the intersection near the cones. Also, there are cars in the parking lots. Tell Euclid to go to each one.

Mission M1: We have received reports that some shipment has been received at the warehouse. Block the road from the warehouse to the barracks, and also inspect the warehouse to identify what has been shipped there. Report when the road is blocked, and report what you observe at the warehouse in a SALUTE report. If you cannot inspect part or all of the shipment, report what you do see and what you cannot inspect.

Note: Since on line activity recognition was not available, we modified this to blocking roads without inspecting the warehouse shipment.

Task: Tell Hamilton to use a cone to block the intersection near the government building. Tell Euclid to use the other cone to block the intersection near the barracks.

Mission M2: Starting at [start] at [time start], your team will conduct a reconnaissance of Area Alpha in order to find Mr. X's vehicle and send back a SALUTE report describing the vehicle and activity around it. Complete the search no later than [time2]. The car may be at one of the following locations [initial position, v1 or v2].

M2 Fragmentary order

We have received reports that Mr. X will be going to the communication center. You should accomplish everything in your original mission, but also block roads leading to the communications center. If you see a traffic cone in front of the communication center, search near the barracks for Mr. X.

Note: We wanted to test modifying the mission with a fragmentary order but decided to leave out the road blocking as we were running out of time and needed a short mission.

Task: Mr. X's vehicle is a car or truck. Tell Hamilton to go to the cars in the parking lot by the government center. Tell Euclid to go to the truck between the barracks and the hospital. FRAGO -- We found Mr. X in the parking lot. Tell Euclid to help Hamilton.

The E and M preceding the numbers categorized missions into "easy" and "medium," and the site of Alpha or Bravo was appended to the names so that missions were recorded as, for example, "E2B" or "M1A" for the second easy mission in Bravo and first medium mission in Alpha.

Appendix B. Evaluation Guide[*]

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

Below is the evaluation guide provided to evaluators for the experiment. Each evaluator had a position or role, and was instructed to record any "steps" of the run that occurred relevant to their role. Steps do not overlap exactly with sub-tasks or actions, as an action or sub-task may be observed by several people, with each person looking at part of it. Interpret the word "step" as a variable that could relate to action or sub-task depending on the role of the observer. The evaluation guide also contains the list of general sub-task grading guides from which the graded actions for research teams were developed. For clarity, teams were not graded based on these, but these were used to choose which actions would be observed to grade teams, with the actions for the MIT team listed in Table 2.

Evaluator and Robot Supervisor Guide

This guide provides general instructions for robotic supervisors and for experiment evaluators. The purpose of this experiment is to identify potential applications of fundamental research conducted under DCIST. There are no planned replications for the experiment, as there are no variables to be estimated in the experimental plan. Instead, evaluation consists of the documentation of the robotic team's ability to perform a series of steps to complete a task, and observations about how it was performed. For those steps attempted, evaluators will record the success or failure of the step, making notes to resolve any ambiguity which might arise after the experiment. As you evaluate, keep in mind you are collecting evidence as to potential applications of research, and document with that in mind.

Evaluator guidance

Each planned run is designed so that an expected sequence of steps will be taken, allowing the robotic team to provide evidence as to how their research can be applied to the task by executing that step. Evaluation sheets show the expected sequence of steps in a diagram, but deviating from this flow is not a failure on behalf of the team. Record success or failure at each step you are observing, making notes as needed when there is ambiguity, multiple attempts at a step, or unexpected steps. Records might include

- Success / failure
- Number of attempts before success
- Measure or notes describing partial success
- Behavior of the robots during the attempt
- Conditions on site you suspect might have influenced robot behavior

After the experiment, write down explanations for your notes while your memory is fresh, sending the explanations and (digital or physical) to the lead for your site.

Record start time when the robots cross the start line, and stop time when the run is finished, as well as the time of events you are noting so they can be later compared to ROS bags.

When you are observing part of a task, such as identifying a car, you are free to interact with researchers or others to gather information as long as you do not interfere with the experiment. The section on sub-task elements lists things that may occur as part of successfully completing part of a task. This is a guide for you to use when determining what has gone wrong if the team is failing. It is not a grading rubric. Use it if it is helpful, but remember that your goal is to record whether the team you are observing successfully completed the sub-task you are observing, and to record conditions of the environment that might have contributed to that, including observations of robot behavior. Use the lists of sub-task elements only to the extent it helps you do that.

Robot supervisor guidance

Robot supervisors are responsible for robots, for example their operation and safety. They should be familiar with the expected behavior of the robot, and able to recognize unusual or potentially unsafe behavior. During a run of the experiment, they have the following responsibilities:

Intervene if the robot is doing something that is unsafe and document the reason for intervention, along with robot and time if possible.

Note unusual or unexpected behaviors of the robot, identifying the time, robot, and observation if possible.

Ensure ROS bags are recorded and labeled to reflect vignette (Easy 1) area (alpha/bravo) version, and attempt (denoting the number of times this same run has been made).

Robot supervisors should make records in whatever format works for them, but collect the records at a team level and send them by team to site leads in some digital format.

Sub-task element lists

Below are subtasks (in italics) we expect to be done during the experiment and a list of elements of those sub-tasks (in a table) that we expect to be present. These lists may help you identify what is succeeding or failing during the experiment.

Table B-1. Mapping of an area.

Relevant navigation terrain classified when visible (e.g. roads or traversable / not traversable)
A path of a certain terrain class between mission relevant points should exist in real life IFF it exists on the map.
Relevant objects, like structures or vehicles, should be mapped in positions so that they will be encountered in the same order during driving as would be expected from the map.
Vehicles and their color should be mapped.
Buildings / structures / shelter should be mapped, including structures relevant to the mission.
People do not need to be mapped, but note if they are.
Obvious metric discrepancies in location should be noted, but metric comparisons should not be needed to determine this. What counts as obvious is determined by qualitative judgement by the evaluator.

Table B-2. Initial planning and starting the mission.

Each robot involved in the mission receives an initial map (if needed)
Each robot identifies their position in the map
Robots receive the mission or instructions relevant to their part in it
Search of the area is initiated, planned with the intended area of operations
Each robot is tracking its own location
Robots share state information as needed and map information as needed
Platform traversable routes are identified
Mapped position of Mr. X vehicle is identified as relevant
Search plan includes visiting the mapped vehicle

Table B-3. Search of an area.

Robots moving in the area correctly identifies relevant objects (see below)
Structures mentioned in the mission are identified via symbol or QR code.
Cones used for blocking roads identified (if mission relevant)
Vehicles of relevant color identified
People identified (if searching for a person)
Pelican cases identified (if part of mission)
Vehicles matching Mr. X's description are inspected, with robot moving around the vehicle to look for the identifying symbol
If all objects in the area identified as not the target object or a mission relevant object, the search of the area stops but the mission continues.

Table B-4. Identify target vehicle.

Target vehicle identified as needing inspection
Robot moves around the vehicle to the part with the identifying symbol
Robot orients sensor to symbol
Robot or operator identifies car by symbol
This and other robots stop search for vehicle and record/report its position (after communicating)

Table B-5. Identify person of interest.

Person identified as a person
Person identified as having distinctive vest
Person identified as the person of interest based on vest
This and other robots stop search for person and record/report its position (after communicating)

Table B-6. Send SALUTE report.

Report contains size description as explicit (1 person, 1 vehicle) or implicit (found Mr. X car or Mr. X).
Activity includes activity relevant to the situation if situation is provided, and may include other activity. If size is implicit, activity should record other instances of the object or person.
Location of the target is reported, by GPS, map region, or other method interpretable to operators.
Robot observing this is identified
Team reports time report is sent or time of observation
Report contains relevant information and no irrelevant information (brevity)
Report contains accurate information and no incorrect information

Table B-7. FRAGO received.

All robots receive FRAGO or instructions based on it
Team adapts plan after receiving FRAGO
New plan includes activity related to relevant locations in original order and FRAGO
Plan does not include activity related to non-relevant locations

Table B-8. Inspect (use) object of a type in an area.

A search of the area is conducted
Objects of correct type (cones, pelican cases) are identified, including multiple instances.
Correct action is taken with respect to object (cone grasped, contents of case identified)

Table B-9. Actions at a roadblock.

Road is identified as being blocked
Roadblock is communicated via message or map update
Alternate routes avoiding the roadblock are taken

Table B-10. Blocking a road.

A region or location on road is identified as relevant area to position blocker
If an object (cone) is to be used as blocker, an instance of the object is identified to be used
A plan is made to retrieve the object instance and position it as blocker
The robot retrieves the instance of the object
The robot moves the object (or itself) to the blocking position and places object or itself there
The robot reports when the blocker (or self) is in place

Appendix C. Experiment Aids

Each run had a setup guide, like the one below, and an evaluation sheet. The setup guide lists roles to be played by U.S. Army Combat Capabilities Development Command (DEVCOM) Army Research Laboratory (ARL) employees, in this case Massachusetts Institute of Technology (MIT) command center (CMD), car observer (Car), and roving observer (Roving). Permitted activity, referred to below, was a list of scenario-relevant things that could be going on.

E1 MIT Execution

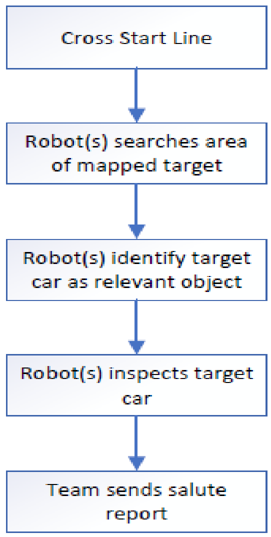
1) MIT Car drives the car to E1-mapped position and places the symbol so the car is visible from the street but the symbol is not. Optionally, stage any permitted activity.
2) MIT CMD checks the mapping of an area to ensure that relevant structures, vehicles, and terrain to the car are mapped to be order-wise and topologically correct for terrain between key points, then MIT CMD directs the team to send the Operations Order.
3) MIT Roving checks initial planning and the start of the mission.
4) MIT Car observes the search around the car and inspection of the car.
5) MIT CMD verifies the car was correctly identified through the Size Activity Location Uniform Time Equipment (SALUTE) report.
6) MIT CMD reads/sends activity to MIT Car for verification.

Each run also had an evaluation guide sheet, like the one below, intended to assist evaluators. The evaluation sheets and setup guides were generally not needed, and planning time might have been saved with a different approach to setup and evaluation.

Evaluator Name Site Alpha/Bravo Observation Point

Start time Time at observation point Stop time

Easy 1 expected flow MIT



Expected subtasks

(select which you observe)

Mapping of an area

Initial planning and starting the mission

Search of an area

Identify target vehicle

Send SALUTE report (see Appendix A)

Notes:

Checklist

Record Robot Operating System bags as robot_easy_1_area_attempt	
Record emergency stops and interventions with reason for intervening	
Mission files and commands provided to robot saved	
Interventions noted by time, robot (if not all) and action taken	
Recorded “unusual” behavior, robot, and time	
Recorded communications sent by and to robot if not in Rosbag	

List of Symbols, Abbreviations, and Acronyms

AI	Artificial Intelligence
ARL	Army Research Laboratory
BMU	Buckner Maintenance Union
BWU	Buckner Warehouse Union
CMD	Command
CRA	Collaborative Research Alliance
DEVCOM	U.S. Army Combat Capabilities Development Command
DCIST	Distributed and Collaborative Intelligent Systems and Technology
FRAGO	Fragmentary Order
GPS	Global Positioning System
LLM	Large Language Model
MIT	Massachusetts Institute of Technology
OPORD	Operations Order
PDDL	Planning Domain Definition Language
ROS	Robot Operating System
SALUTE	Size Activity Location Uniform Time Equipment

1 (PDF)	DEFENSE TECHNICAL INFORMATION CTR DTIC OCA
1 (PDF)	DEVCOM ARL FCDD RLB CB TECH LIB