

Analysis of Human-Robot Interaction at the DARPA Robotics Challenge Trials

Holly A. Yanco¹, Adam Norton², Willard Ober³, David Shane³, Anna Skinner⁴, and Jack Vice⁴

¹Computer Science Department, University of Massachusetts Lowell, One University Avenue, Lowell, MA 01854, holly@cs.uml.edu

²New England Robotics Validation and Experimentation (NERVE) Center, University of Massachusetts Lowell, 1001 Pawtucket Blvd, Lowell, MA 01854, anorton@cs.uml.edu

³Boston Engineering Corporation, 300 Bear Hill Road, Waltham, MA 02451, {wober, dshane}@boston-engineering.com

⁴AnthroTronix, 8737 Colesville Rd, Silver Spring, MD 20910, {askinner, jvice}@atinc.com

Abstract

In December 2013, the Defense Advanced Research Projects Agency (DARPA) Robotics Challenge (DRC) Trials were held in Homestead, Florida. The DRC Trials were designed to test the capabilities of humanoid robots in disaster response scenarios with degraded communications. Each team created their own interaction method to control their robot, either the Boston Dynamics Atlas robot or a robot built by the team themselves. Of the fifteen competing teams, eight participated in our study of human-robot interaction (HRI). We observed the participating teams from the field (with the robot) and in the control room (with the operators), noting many performance metrics, such as critical incidents and utterances, and categorizing their interaction methods according to number of operators, control methods, and interface automation. We decomposed each task into a series of subtasks, different from the DRC Trials official subtasks for points, to gain a better understanding of each team's performance in varying complexities of mobility and manipulation. Each team's interaction methods have been compared to their performance and correlations have been analyzed to understand why some teams ranked higher than others. We discuss lessons learned from this study, and have found in general that the guidelines for human-robot interaction for unmanned ground vehicles still hold true: more sensor fusion, fewer operators, and more automation lead to better performance.

1. Introduction

The Defense Advanced Research Projects Agency (DARPA) began conducting robotics challenges in 2004 and 2005 with the DARPA Grand Challenge, an autonomous vehicle race across the desert in California. The idea was to offer a large prize, as well as a considerable amount of publicity, for the winning team in a 150 mile autonomous unmanned ground vehicle (UGV) off-road race. This challenge evolved into one focused on an urban landscape in the 2007 DARPA Urban Challenge. As ground vehicle research migrated to commercial and consumer companies, and in light of the nuclear power plant disaster at the Fukushima Daiichi plant in Japan after the earthquake and tsunami in March 2011, DARPA developed a new challenge in 2012, the DARPA Robotics Challenge (DRC).

For the DRC, tasks were modeled within the context of Urban Search and Rescue (USAR) and industrial disaster response task domains, involving a set of mobility and manipulation hazardous duty, real world anthropomorphic tasks conducted using a combination of interface automation and teleoperation. The DRC is a multi-year challenge designed to radically improve the state of the art in rescue robotics, with the goal of having robot systems that could have assisted during the disaster at Fukushima Daiichi. Beginning in 2012 with the Virtual Robotics Challenge (VRC), and then moving to physical hardware and human-engineered physical environments in the Trials in December 2013 and the Finals in June 2015, the DRC provides a standardized testing environment in which to assess current capabilities and compare across robotics development teams. Some teams were provided with a standardized robotic platform: the Atlas humanoid robot, developed by Boston Dynamics [BDI 2014], with 28 degrees of freedom, designed for mobility in outdoor environments and uneven terrain as well as for manipulating objects such as valves and tools.

As part of contract W31P4Q-13-C-0136 with Boston Engineering Corporation and the University of Massachusetts Lowell and contract W31P4Q-13C-0196 with AnthroTronix and the University of Southern California sponsored by the DARPA Defense Sciences Office, we were invited to study the DRC Trials, with the goal of identifying areas for improvement that would lead to better HRI design and overall robot performance in the DRC Finals. Given that each competing team had to design their own interaction method, our team used the DRC Trials as a baseline for gathering data about a very future-forward application of robots for disaster response. The ultimate goal of human-robot collaborative interaction is to exploit the strengths and capabilities of both the human and robot team members while compensating for the weaknesses and shortcomings of each, in order to maximize effectiveness of the team in completing mission goals safely and effectively. Within such contexts, HRI designers must consider aspects of the control interface that may impede or negate the human's strengths and capabilities.

Despite early artificial intelligence (AI) predictions that machines would be capable of performing any tasks that humans can perform by this time, the field is far from achieving this goal; even basic tasks continue to require supervisory control from one or more operators, and complex tasks such as those involved in USAR domains often require continuous direct control, often by multiple operators. Therefore, the ability of robots to perform functions effectively depends, in part, on the ability of humans to control or interact with them effectively, constraining the robot's performance to the operator's skill and the design of the interface [Fong, Thorpe, and Baur 2002]. In order for a human-robot interface to be effective in a dynamic, hazardous environment, the interface must be efficient, intuitive, unobtrusive, intelligent, and, ideally, adaptable. Much research has been dedicated to the development of advanced robotic systems with increased autonomy, but comparatively minimal research has been dedicated to the development of scientifically validated control interfaces. Interface design has continued to improve; however, it remains more of an art than a science. Interface designers must consider multiple control paradigms based not only on the characteristics of the robot, but of the user as well.

2. Background and Related Work

Given that hardware for humanoid robots has only been in active development for about two decades, the topic of HRI design for humanoid robots is a fairly new one. One of the most developed interfaces for a humanoid robot was created to teleoperate NASA's Robonaut and Robonaut 2 robots [Diffler et al. 2003; Fong et al. 2013]. HRI for the Robonaut robots requires a

highly skilled and trained operator, in an environment that allows for simulation (e.g., NASA's full sized mock up of the International Space Station) and mission planning. In contrast, after the tsunami hit the Fukushima Daiichi plant, a rapid response was needed in an environment that no longer matched the as-built plans. Additionally, the people best qualified to understand the situation inside the plant, the nuclear engineers, were not skilled robot operators. HRI design for another humanoid system used two joysticks to control the robot [Sian et al. 2002], which also required training on the system's movement.

When developing HRI for robots intended for USAR, there is a need to create systems that can be used by first responders as easily as possible, with the focus on ease of learnability (e.g., [Micire 2010]). The DRC Trials offered the first opportunity to observe a large number of HRI approaches for USAR with humanoid robots.

Our study leveraged the experience of our multi-disciplinary team in conducting HRI evaluations, including previous HRI studies conducted within the context of robotic competitions. We also reviewed relevant literature and field studies within this domain, using best practices and lessons learned to guide the development of metrics and data collection techniques, as well as data distillation, analysis, and interpretation methods to facilitate thoroughness in data collection and evaluation methodologies without interrupting task workflow. These studies included evaluations of the American Association for Artificial Intelligence (AAAI; now called the Association for the Advancement of Artificial Intelligence) and RoboCup Robot Rescue Competitions [Scholtz et al. 2004; Yanco, Drury & Scholtz 2004; Yanco and Drury 2007], as well as USAR disaster response and field exercise evaluations [Murphy and Burke 2005; Nourbakhsh et al. 2005]. Adams [2002] describes how interface methods can have direct effects on situation awareness and workload levels. Scholtz [2002] developed a set of evaluation criteria for assessing overall human-intelligent system interaction. Additionally, Nourbakhsh et al. [2005] provided promising solutions to the vast complexity of USAR HRI such as multi-agent systems and simulation and control interfaces that incorporate the various levels of robot autonomy.

These studies, along with others, guided us in using best practices and lessons learned to guide the development of our metrics and data collection techniques, as well as methods for data distillation, analysis, and ensuring thoroughness in data collection and evaluation methodologies without interrupting task workflow. That said, the DRC Trials was easily the richest and most advanced display of humanoid robots in competition, providing a unique evaluation opportunity.

In a multi-year study (2002-2004) examining HRI issues within the context of the AAI/RoboCup Robot Rescue Competition, Yanco and Drury [2007] evaluated the impact of various HRI approaches on competition performance. The outcomes of this study led to the identification of five primary guidelines applicable to the design of HRI within the USAR domain:

1. Utilize a single monitor for the interface.
2. Avoid small video windows on the interface.
3. Avoid window occlusion.
4. Use one robot to view another when more than one robot is available.
5. Design for the intended user, not the developer.

These guidelines were established for the development of HRI interfaces for a single operator, whereas all of the observed DRC Trials teams employed multiple operators (counting both passive and active operators), working together by viewing multiple interfaces simultaneously.

Yanco and Drury [2007] also highlighted the importance of SA and operator's SA strategies within the context of USAR robotic control tasks, as well as the inherent limitations in assessing SA both in real-time and via post hoc interviews and analyses. SA issues are especially problematic within dynamic HRI scenarios in which changes may be simultaneously occurring in the operator, the robot, and the environment. The 2013 DRC Trials task environments were static; however, future competitions are anticipated to involve more dynamic environments, and certainly transfer to real world scenarios will involve task performance within unpredictable and hazardous changing environments. Therefore, the evaluation of HRI must take into account all factors impacting the SA of the human, the SA of the robot, and the resulting shared SA.

3. DRC Trials Tasks

The DRC Trials were designed to test the capabilities of teams with individual, self-contained tasks whose layouts were known in advance and did not change during the competition [DRC 2014]. To simulate difficulties with communications networks during an actual disaster, DARPA alternated periods of high and low bandwidth for data sent between the robot and the control room, lasting one minute each. "Good comms" had a data rate of 1 MB/second in each direction and a delay of 100 ms round trip (50 ms each direction). "Bad comms" had a data rate of 1 KB/second in each direction and a delay of 1,000 ms round trip (500 ms each direction).

Below are brief descriptions of the seven tasks we observed at the DRC Trials (we did not observe the Vehicle task). Images of each task can be found in Table 5. Based on overall task completion points by all competing teams at the DRC Trials, the tasks were ranked by DARPA in terms of difficulty as Valve (easiest); Terrain and Hose (easier); Door, Debris, Wall and Ladder (harder); Vehicle (hardest) [DARPA 2014]. The DRC tasks mainly fall into two of the seven missions for which USAR UGVs have been used in the past, direct intervention and rubble removal and clearing [Murphy 2014].

The **Terrain** task consisted of three zones, each of which had start and end lines which had to be crossed in order to be considered complete. All eight teams participating in our study were observed on the Terrain task; however, our observation of Team B was unable to be used for a reason we will not explain here to preserve anonymity. This task is almost entirely a mobility task, aside from manipulators being used for balance, involving obstacles that can cause robots to slip, trip, and fall.

The **Ladder** task consisted of three zones, each of which had a number of steps (1st step, 4th step, and 9th step/landing platform, respectively) that the robot had to stand on with both feet to be considered complete. Unlike other tasks, some teams performed this task with their robot facing backwards. Four of our participating teams were observed on the Ladder task. This task combined both mobility and manipulation, although some teams did not utilize their robot's manipulators to aid in climbing the ladder.

The **Debris** task consisted of two sets of debris that needed to be removed, a truss that can optionally be removed, and a doorway which must be traversed through for the task to be completed. Four of our participating teams were observed on the Debris task. This task was mainly focused on manipulation, although mobility was used to orient the robot's body in order to remove debris, and to traverse through the doorway and around the truss, if applicable.

The **Door** task consisted of three doors that had to be opened and fully moved through to be

considered complete. Each door had a lever-style handle that had to be turned in order for the door to open. All eight teams participating in our study were observed on the Door task. Both mobility and manipulation were equally important to this task, requiring teams to first use them separately during the first two doors. Teams had to combine them during the third door (which was weighted) in order to hold the door open while traversing through it. Due to wind conditions, some teams needed to hold the earlier doors as well to prevent them from blowing closed.

The **Wall** task consisted of a wall segment that had to be cut and then removed. The robot needed to first walk over to a shelf that contained a hand drill, pick up the hand drill, and turn it on. The robot then had to carry the drill over to the wall and use it to make three cuts in the wall (forming a triangle), then remove the cut piece. Five of our participating teams were observed on the Wall task. This task was mainly focused on manipulation, although mobility was used to walk to the shelf, to the wall, and to orient the robot while picking up the drill and cut into the wall.

The **Valve** task consisted of three valves whose wheels or levers needed to be rotated to close them. The valves could be closed in any order. All eight teams participating in our study were observed on the Valve task. This task was mainly focused on manipulation, although mobility was required to walk to each valve. Some teams opted to use a single hand to close valves while others used two arms/hands in tandem.

The **Hose** task consisted of a hose on a reel that had to be grasped, unreeled, carried to a wye, and attached to the wye. Six of our participating teams were observed on the Hose task. This task was mainly focused on manipulation, although mobility was used to walk to the hose, to the wye, and to orient the robot while picking up the hose, unreeling the hose, and attaching the hose to the wye.

4. Methodology

Our study was approved by the Institutional Review Board (IRB) at the University of Massachusetts Lowell. Participants were invited to join the study via email approximately two weeks before the DRC Trials. Teams who decided to participate completed an informed consent form and a pre-DRC Trials questionnaire about the design of their robot system. Eight of the fifteen teams who competed in the DRC Trials elected to participate in our study.

We are required by our IRB to anonymize the results discussed in this paper. Therefore, we will label the eight participating teams as Teams A through H. To allow for some meaning in these labels, we have sorted the participating teams by their overall scores. To prevent identification based upon publicly available data showing the correspondence of team names and overall scores, we do not report the overall point total for any teams in this paper. Additionally, in cases of tie scores, we have randomly put one team in front of the other without identifying that there was a tie. Of the eight participating teams, Teams A through D were in the top 8, and Teams E through H were in the bottom 7. We also do not report information about the robot being used by each team, to avoid identifying teams with unique systems.

Our evaluation team consisted of eight people, split into four pairs. Our overall goal was to evaluate the impact of the interface features and the control methods on the effectiveness of the teams. We set an observation schedule for the two days of the event based upon a few constraints:

1. Our first priority was to observe all of the participating teams on three tasks: Valve

(primarily a manipulation task), Terrain (primarily a mobility task), and Door (which combined both manipulation and mobility).

2. Our second priority was to observe all of the participating teams on as many additional tasks as possible. On the practice day, evaluation team members met with team representatives to interview them about their HRI designs and ask additional robot system design questions. In this meeting, we asked teams to tell us which events they thought would be their top three events and to also tell us in which events they were planning not to complete. We found that several of the teams did not plan to participate in the Vehicle task, so we removed that task from the set of tasks we would observe. We started filling the remaining slots by ensuring that we covered each team’s predicted best events. Finally, we filled any free slots after that by aiming to balance the number of observations of each team and of each task.

The matrix of teams and tasks, with an X denoting that the pair was evaluated in our study, is shown in Table 1.

	Terrain	Ladder	Debris	Door	Wall	Valve	Hose	Total
Team A	X		X	X	X	X	X	6
Team B	---	X	X	X		X	X	6
Team C	X			X	X	X	X	5
Team D	X	X		X	X	X	X	6
Team E	X	X		X	X	X	X	6
Team F	X	X	X	X		X	X	6
Team G	X		X	X		X		4
Team H	X			X	X	X		4
Total	8	4	4	8	4	8	6	43

Table 1. Observation coverage across tasks for each team participating in our study. An X indicates that we observed the team on that task and included the data in our analysis. A dash means that we observed the team, but we were not able to include the data in our analysis (the reason can not be stated for purposes of anonymization).

Our team designed data collection sheets to be used for recording observations on the field and in the control rooms during the trials. During the task runs, the field observer recorded robot movements and critical incidents, as well as official time and scoring information. At the same time, the control room observer took handwritten notes to record team dialog and information about the use of the interface. The control room observer also sketched the arrangement of the operators, the displays, and interaction methods before the run started. We noted the number of active operators (people directly controlling the robot) and passive operators (people offering advice or watching over the shoulders of the active operators).

From the data we recorded, we extracted several different metrics, including successful subtask attempts, utterance coding, and percentage of aggregate critical incidents per team. We compared these metrics to aspects of each competing teams’ interaction methods (see Section 6) to find correlations, including subtask types vs. control methods, task-oriented utterances vs. robot-oriented utterances, etc. Our methods for analyzing the data collected are described in Section 5.

5. Analysis Methods

Our analysis relied primarily on the notes taken by our field and control room observers. In a few instances, we used the DRC Trials video on YouTube [DRC Videos 2014] to clarify our data.

5.1 Field and Control Room Notes

After the DRC Trials, each handwritten set of observation sheets was typed into its own spreadsheet by the observer who wrote the notes. After that step, each pair of digitized control room and field data sheets was combined to form a single sheet for each task per team. This combination allowed for links to be drawn between control room observations and field observations. We analyzed these combined sheets to assess the situation awareness of the team; for example, the combined sheets would allow us to determine if the control room knew that a critical incident had happened on the field. Table 2 shows an example of the combined notes for Team F on the Terrain task. From the field, it appeared that the robot just lost its balance, causing it to fall. By using the side-by-side presentation, we can see that the operator forgot to click a menu item on the interface that was used to switch between walking states, causing the robot to fall.

Field	Code	Control room
Body shift forward, head up, body shift back, fall = INTERVENTION	FALL	Op2: "What happened?" Op1: "I didn't click it" Op2: "Balance didn't go on." Op2: "We can do an intervention."

Table 2. Combined field and control room data sheet example for Team F on the Terrain task.

5.2 Critical Incidents

Based upon our knowledge of the tasks at the DRC Trials and the problems that robots can encounter, we identified the critical incidents that might occur before the competition, defined as follows:

- **TIP (T):** The robot begins to lean noticeably to one side unintentionally. (Note that the normal side to side movement for walking was not coded as a TIP critical incident.) The critical incident was coded even when the robot was able to recover.
- **HIT (H):** Part of the robot's body/limb hits part of the apparatus unintentionally.
- **TRIP (Tr):** The robot's foot or leg snags part of the apparatus, causing it to fall or tip.
- **MISS (M):** The robot misses while attempting to grasp an object or places a footfall perceived to be incorrect.
- **STUCK (St):** Part of the robot's body/limb is stuck on or in part of the apparatus, potentially causing a trip, fall, and/or intervention. No instances of this type were observed in our study.
- **SLIP (S):** The robot's limb slips off of part of the apparatus.
- **DROP (D):** An object the robot was carrying is dropped unintentionally.
- **FALL (F):** The robot falls and the belay is triggered, potentially followed by an intervention.
- **RESET (R):** When an intervention was called that did not correspond to another critical incident (e.g. some teams called interventions to check sensor readings).

A total of 77 critical incidents were observed as part of this study, of which 33 were falls and/or resets (called out separately due to the fact that when a fall and/or reset occurs an intervention or end of run is called; a team is charged 5 minutes for an intervention, which is taken away from their 30 minute run time) and 44 were other critical incidents. Table 3 outlines the critical incidents observed per team and per task, listing the types of critical incidents observed.

	Terrain	Ladder	Debris	Door	Wall	Valve	Hose	Total falls and/or resets	Total other critical incidents
Team A	---		---	S	---	F	---	1	1
Team B		F	H	F, T, H		M	S, D	2	6
Team C	F, F			S, S	M	---	M, M, M, M	2	7
Team D	F, F, H	F		R, F	D	H	H, H, H, H, H, D	5	9
Team E	F, F, M	R, F		F, F, H, H	---	S	---	6	4
Team F	F	F	F, F			R, F, F	---	7	0
Team G	F		H, H, H, M, M	F, S		F		3	6
Team H	R, R			S, S, S	R, R, R	R, R, H, H, H, H, H, S, S, S		7	11
Total falls and/or resets	10	5	2	6	3	7	0	33	
Total other critical incidents	2	0	6	11	2	11	12		44

Table 3. All critical incidents observed per team per task. Dashes indicate that no critical incidents were observed for a team on that task. A gray cell indicates that the team was not observed on that task or, in the case of Team B on Terrain, the data could not be included.

5.3 Subtasks for HRI Evaluation

Many teams did not fully complete tasks, generally earning only 1 or 2 points, by completing 1 or 2 DRC-defined subtasks. However, most of the DRC-defined subtasks involved many smaller actions, each of which could prompt a change in each team’s interaction method. For this reason, we broke down each DRC Trials task into finer grain subtasks than those needed to score points. For instance, the Debris task was decomposed into 13 subtasks: traversing from the starting point to the debris pile, removing each of the yellow debris pieces (5 individual subtasks), removing each of the orange debris pieces (5 individual sub-tasks), removing the truss (if applicable), and traversing from the debris pile through the doorway.

While some DRC-defined subtasks were comprised of entirely mobility or manipulation (e.g., each subtask on Terrain was mobility-specific), others required both to be accomplished (e.g., in the Hose task, traversing to the wye with the hose requires first walking to the hose, grasping it, unreeling it, and then traversing with the hose in the hand to be accomplished). We defined four subtask functions: unobstructed traverse (UT), obstructed traverse (OT), manipulation (M), or second order manipulation (SOM), referring to the manipulation of a tool within the environment. After breaking down the tasks into smaller subtasks, each subtask was then categorized as one of these functions. With these codings, each team’s timing per task can be summed to report the total amount of time spent performing subtasks in each function. Table 4 shows an example of subtask timing comparisons between Team B and Team E on the Hose task. Table 5 outlines all of the subtasks and their categorizations for each task. Across all of the tasks, there are 27 mobility subtasks (11 UT and 16 OT) and 31 manipulation subtasks (7 FOM and 24 SOM).

	Team B	Team E
Traverse to hose complete	N/A	2
Grasping hose or nozzle complete	1	2
Unreeling hose complete	6	2
Traverse with hose to wye complete	3	2
Point 1	10	8
Hose nozzle touching wye complete	7	9
Point 2	17	17
Hose nozzle rotated to attach to wye complete	---	---
Releasing nozzle attached to wye complete	---	---
Point 3	---	---

Table 4. Teams B and E on the Hose task. Note: a marking of N/A means the subtask’s timing was not discernable from our observation notes. A marking of “---” means the team did not attempt that subtask. The measures in this table refer to the number of minutes it took each team to complete that subtask.

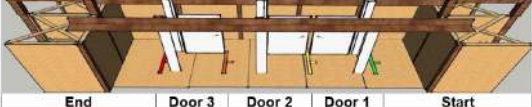
5.4 Utterance Coding

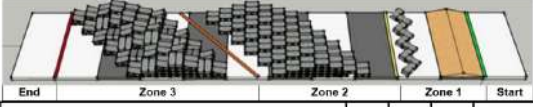
The majority of handwritten notes taken in each team’s garage during tasks included dialog between team members. (No audio recordings were made in the control rooms.) No dialog was noted for Team C on Valve or Team F on Ladder. The control room observations for Team F on Hose were taken at the incorrect operator station during a team practice, so we have no recorded utterances for Team F on Hose. In our analysis of coded utterances, we omitted these three observations. The data set provided a total of 1397 utterances over the 40 observed team/task pairs with usable utterances. Utterances were coded in five categories; the categories were influenced by the dialog recorded, rather than predefined, using the grounded theory method [Glaser and Strauss 1967]. The five main categories coded were subject, situation awareness, type, classification and emotion.

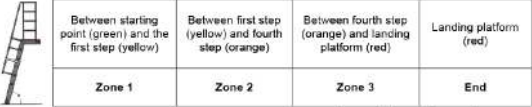
The **subject** category was used to identify the main topic, using the following classes:

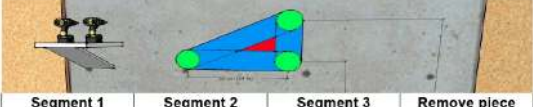
- **Comms:** Utterance referred to the state or change of the communication line between the team’s interaction method and the robot.
- **Time:** Utterance referred to the amount of time remaining in the run, an intervention, or for a robot action to execute.
- **Interface:** Utterance referred to a data display within the interface or control method.
- **Robot: General:** Utterance referred to the robot, but not the robot’s arms/hands or legs/feet.
- **Robot: Legs/Feet:** Utterance referred to the robot’s legs or feet, usually about mobility.
- **Robot: Arms/Hands:** Utterance referred to the robot’s arms or hands, usually about manipulation.
- **Obstacle:** Utterance referred to an obstacle that the robot can potentially bump into or trip over (e.g. cinder block steps, doorway frame, etc.) or catch the robot’s hand or arm on (e.g., wall next to the debris, pipes on which the valves were attached).
- **Tool:** Utterance referred to an object in the environment that was meant to be manipulated (e.g. piece of debris, door handle, drill, valve wheel, hose nozzle, etc.).
- **Patter:** Used for encouragement or frustration (e.g. “Good job!” or “Stupid...”).
- **Action Plan:** Utterances that referred to a plan to do something that were not coded into any of the other categories.

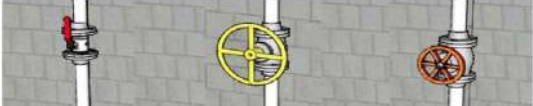
More than one coding could be used for each utterance; for example, if an utterance mentioned opening the robot’s hand to grip the drill, it coded as Robot: Arms/Hands and Tool (drill).

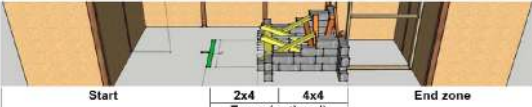
				
Door	UT	OT	M	SOM
Traverse ground to push door	X			
Push door open				X
Travel completely through push door		X		X
Pull door open				X
Travel completely through pull door		X		X
Weighted pull door open				X
Travel completely through weighted pull door		X		X

				
Terrain	UT	OT	M	SOM
Traverse ramps		X		
Traverse hurdle		X		
Traverse block steps		X		
Traverse slanted block steps		X		
Walk to finish line (if applicable)	X			

				
Ladder	UT	OT	M	SOM
Traverse ground to ladder	X			
Standing on first step		X		
Standing on second step		X		
Standing on third step		X		
Standing on fourth step		X		
Standing on fifth step		X		
Standing on sixth step		X		
Standing on seventh step		X		
Standing on eighth step		X		
Standing on platform		X		

				
Wall	UT	OT	M	SOM
Traverse ground to shelf	X			
Grasp grill from shelf and pick up			X	
Cut first segment in wall				X
Cut second segment in wall				X
Cut third segment in wall				X
Remove wall piece			X	

				
Valve	UT	OT	M	SOM
Aligning with ball valve (if applicable)	X			
Rotation of ball valve			X	
Traverse ground and align with large rotary valve	X			
Rotation of large rotary valve			X	
Traverse ground and align with mid-size rotary valve	X			
Rotation of mid-size rotary valve			X	

				
Debris	UT	OT	M	SOM
Traverse ground to debris pile	X			
Remove 1st piece of yellow debris				X
Remove 2nd piece of yellow debris				X
Remove 3rd piece of yellow debris				X
Remove 4th piece of yellow debris				X
Remove 5th piece of yellow debris				X
Remove 1st piece of orange debris				X
Remove 2nd piece of orange debris				X
Remove 3rd piece of orange debris				X
Remove 4th piece of orange debris				X
Remove 5th piece of orange debris				X
Remove truss complete (if applicable)				X
Traverse ground through doorway	X			

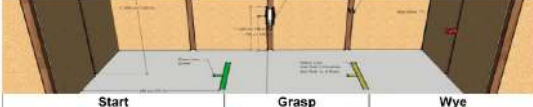
				
Hose	UT	OT	M	SOM
Traverse to hose complete	X			
Grasping hose or nozzle complete			X	
Unreeling hose complete			X	
Traverse with hose to wye complete	X			
Hose nozzle touching wye complete				X
Hose nozzle rotated to attach to wye complete				X
Releasing nozzle attached to wye complete				X

Table 5. Outline of each task's breakdown into subtasks and each subtask's corresponding function(s).

The **situation awareness (SA)** category was used to code whether or not the utterance indicated an awareness, or lack thereof, of the state of the robot’s limbs, the environment state, etc. Utterances were not coded as SA if they did not refer to it.

- **Good SA:** Used when an utterance referred to correct/positive awareness of the state of the robot, the environment, etc. (e.g. “It is not turning anymore”, referring to the drill bit having stopped moving).
- **Neutral SA:** Used when an utterance referred to awareness of the state of the robot, the environment, etc. (e.g. “Check my grasp” where the speaker is aware that there may be a problem with their manipulator’s grasp).
- **Bad SA:** Used when an utterance referred to incorrect/negative awareness of the state of the robot, the environment, etc. (e.g. “I can’t tell if the drill is spinning”).

The **type** category was used to code the utterance’s purpose.

- **Command:** Used when an utterance was instructing another team member to do something.
- **Inquiry:** Used when an utterance was asking a question.
- **Response:** Used when an utterance was responding to an inquiry.
- **Notification:** Used for unprompted statements.

The **classification** category was used to code whether an utterance was positive, negative or neutral. The **emotion** category was used to code if an utterance displayed any stress or encouragement. A majority of the statements were coded as having neither.

Table 6 shows a sample of utterance coding from Team E on the Wall task.

	Subject	SA	Type	Classification	Emotion
“If it’s not turning...”	Tool	Neutral SA	Notification	Neutral	None
“I thought you guys said it was.”	Tool	Bad SA	Notification	Neutral	None
“Stuck?”	Tool	Neutral SA	Inquiry	Neutral	None
“It is stuck.”	Tool	Good SA	Response	Neutral	None
“4 minutes”	Time	---	Notification	Neutral	None

Table 6. An example of utterance coding from Team E on the Wall task.

Inter-rater reliability was established using Cohen’s Kappa for each of the classification groups. For the subject of the utterance, $\kappa=0.84$ excluding chance ($\kappa=0.88$ if chance was not factored out). For situation awareness, $\kappa=0.63$ excluding chance ($\kappa=0.80$ if chance was not factored out). For utterance type, $\kappa=1.0$ excluding chance ($\kappa=1.0$ if chance was not factored out). For the classification, $\kappa=0.67$ excluding chance ($\kappa=0.88$ if chance was not factored out). Finally, for emotion, $\kappa=0.55$ excluding chance ($\kappa=0.81$ if chance was not factored out). Emotion was the most difficult category to code, given that some observers used exclamation points to indicate tone of voice, while others did not.

6. Team Interaction Methods

Each of the eight teams observed during the DRC Trials custom-designed their human-robot interaction (HRI) methods and interaction tools. Each consisted of a different number of display screens, input methods (although every team used at least one keyboard and mouse), sensor fusion, and autonomy levels, among other factors. This section describes the HRI method, interface technology, and team dynamics for each of the competing teams.

6.1 Interface Displays

Table 7 shows the type of data displays each team’s interaction method used as well as the number of unique instances of each; this number mostly varied on number of simultaneous camera feeds displayed. Table 8 shows the data displays that were used to form a unique instance of sensor fusion in each team’s interaction method. The data displays checked in each column were able to be fused with each other within a single window on a screen, but most had the ability to turn the fusion on and off depending on operator and task preferences. Table 9 shows another unique instance of sensor fusion for each team, if applicable. Four teams had a single instance of unique sensor fusion, while the other four teams had two instances of unique sensor fusion. It is important to note that our distinction for each team’s use of sensor fusion is purely from an operator perspective, meaning the display method for the operator to interpret the data. This measure does not take into account any back end processes used to fuse the data before presenting it to the operator, although we can assume that this occurs at some level in order for it to be represented appropriately.

Note that the top four teams had a second sensor fusion method, reported in Table 10, while the bottom four teams did not. In Section 8, we present further results showing that sensor fusion in the human-robot interface leads to better outcomes across several metrics.

	Camera feeds	Point cloud	3D robot avatar	Sim objects or obstacles	2D distance visualization	2D height maps	Status messages, sensor readings, etc.	Total
Team A	3	1	1	1			1	7
Team B	3	1	1	1			1	7
Team C	2	1	1			2	1	7
Team D	4	1	1				1	7
Team E	4	1	1	1	1		1	9
Team F	1	1	1	1			1	5
Team G	6	1	1				1	9
Team H	1				1		1	3

Table 7. Number of unique data displays per team.

	Single camera view	Foveated vision	Point cloud	3D robot avatar	2D distance visualization	Sim objects or obstacles	Joint sensor readings	Notes
Team A	X	X				X		Left window, center/right screen
Team B	X	X						Different screen than 2
Team C	X		X	X				Viewable on right screen
Team D			X	X				Duplicated throughout interface screens
Team E	X	X	X	X		X		Viewable throughout interface screens with different data displays turned on or off
Team F			X	X		X		Viewable throughout interface screens with different data displays turned on or off
Team G			X	X				Viewable throughout interface screens
Team H	X				X			Viewable on right screen

Table 8. Number of combined data displays for one unique instance of sensor fusion per team.

	Single camera view	Foveated vision	Point cloud	3D robot avatar	2D distance visualization	Sim objects or obstacles	Joint sensor readings	Notes
Team A	X		X	X		X		Right window, center/right screen
Team B			X	X		X		Different screen than 1
Team C	X		X					Viewable on right screen in separate window
Team D	X						X	Used to highlight hand in camera views

Table 9. Number of combined data displays for another unique instance of sensor fusion per team. Teams E through H are not listed because they only had one unique instance of sensor fusion.

6.2 Operators, Input Devices, and Screens

The composition of each team’s set of operators in their control room varied greatly. The roles of the operators on each team also varied, but there were some common roles. Most teams had primary operators who were responsible for sending commands to the robot that resulted in robot movement, planning operators who were responsible for programming appropriate plans for the robot to execute, and operators who were focused on strategy or robot status, such as monitoring comms levels. Operators were generally confined to a cluster of display screens, which we refer to as stations. Figure 1 provides an overview image of the composition of all teams’ average operator/screen ratio.

Team A used a single station consisting of three monitors. The left monitor was used to launch the robot’s controller at the start of the run. The center and right screens displayed the same information, with the center used by the operator and the right by a supervisor. During task execution, the

supervisor provided a second set of eyes and aided the operator with strategy and reminders during task execution. Team A throttled down the amount of bandwidth needed to communicate with their robot such that changes in comms strength had no effect on their performance.

Team B used four stations to control their robot. The left station was used by the primary operator to control the robot and to plan robot actions. The center right station was used by a perception operator who was responsible for placing pre-made 3D models of objects into the environment for the primary operator. The right station was used for checking status and communications. The center left station was used for cues and configuration changes during task execution.

Team C used a single station with two displays (left and right). The right screen displayed state information (e.g. network connectivity, command feedback, data message rates) and most other data from their robot. All content on the right screen could be hidden or displayed based on operator preference and task needs. The left screen generally displayed a single camera feed, but was able to display any of the available data displays. To reduce the amount of data sent over the network, point cloud/camera feed data was displayed in grayscale and the point cloud data was discretized to varying levels of detail, depending on the task being performed (e.g. lower resolution for mobility and higher resolution for manipulation).

Team D used four stations to control their robot. The front left and front right stations were used by the two primary operators. Their screens shared the same content, although the bottom two screens and top two screens were switched between stations, (i.e. the front left station's bottom two screens would display the same content as the front right station's top two screens) allowing the two primary operators to see what the other was doing. The back left and back right stations displayed a subset of the data displays from the front left and front right stations to supervisors.

Team E used three stations to control their robot. Every screen at each station was able to display all of the possible data displays. Each one of the center station monitors was primarily dedicated to one of the data displays. The center station was used by the primary operator to plan and perform each task, and was the only station that sent motion commands to the robot. The right station was used for acquiring new and updating existing 3D point clouds and camera images used in the center station. The operator at the right station also had the ability to identify points of interest within the 3D point clouds and camera images for the primary operator. The left station was mainly used for communications and status checks of the robot's onboard computer.

Team F used two interaction method setups. Method 1 used three stations, as did method 2, albeit in a different layout. Throughout their interface, a single camera feed was displayed along with a fusion of a 3D robot avatar and point cloud data. Status messages and sensor readings were also displayed.

Team G used three stations to control their robot. Each of the three stations had two display screens. A game controller was used on the right station to manually place footsteps within the environment for the robot to execute.

Team H used three stations to control their robot. Each of the three stations had one display screen. A gesture recognition device was used in the right station to control the hands and arms of the robot.

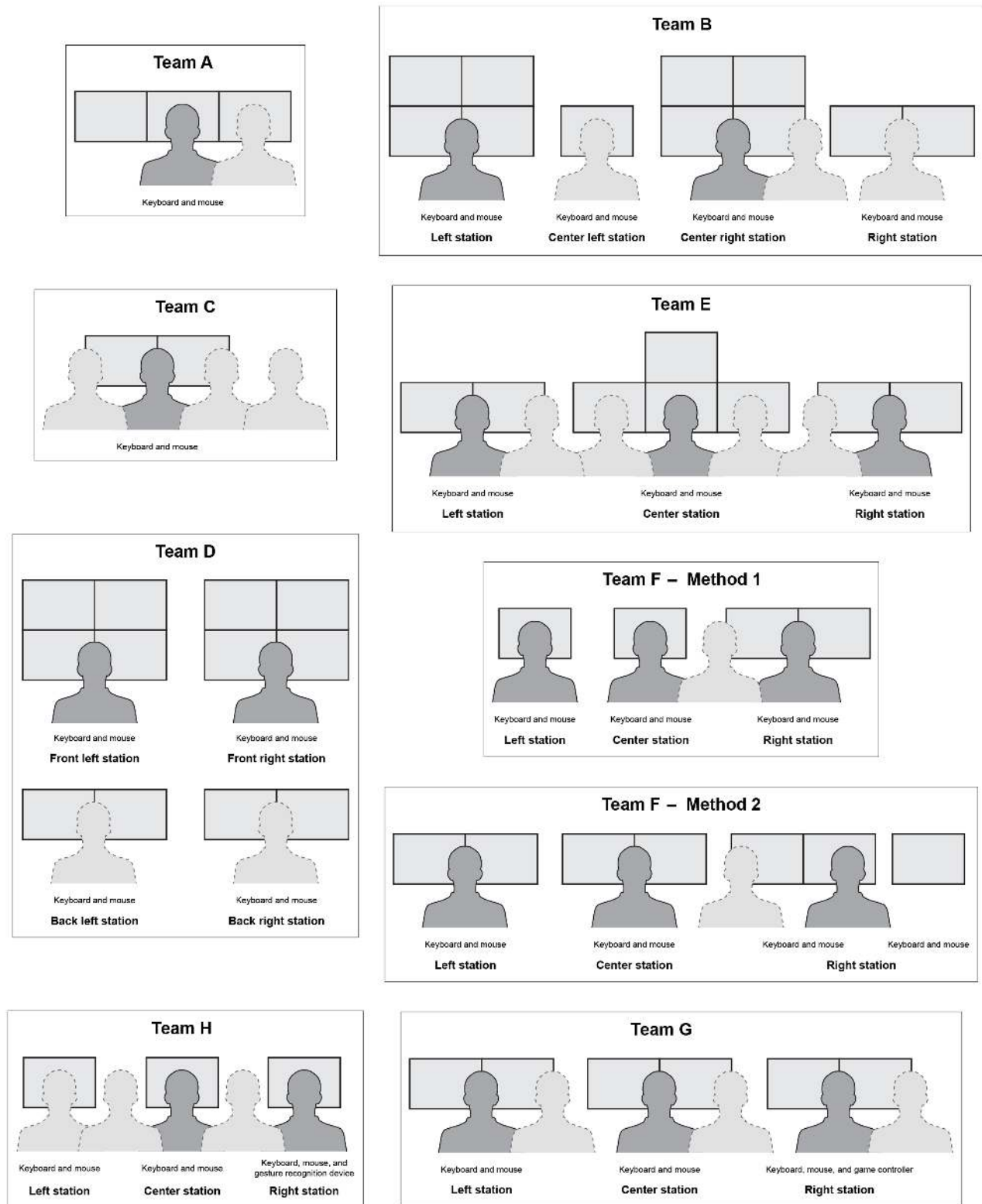


Figure 1. Operator/screen ratios and set-ups for all teams that participated in our study. Active operators are rendered darker and passive operators are rendered lighter.

Table 10 shows the average number of active, passive, and total operators for each team on the observed tasks as well as the total input devices and screens used as part of their interaction method (for Team F the maximum number of input devices and screens was used, given that they had two operating configurations). We defined an active operator as someone controlling an input device for robot control and a passive operator as someone observing task execution, providing feedback on comms, sensor readings, strategic advice, etc. Anyone present in the garage during task execution who was watching or supporting an active operator was considered to be a passive operator.

	Active operators (average)	Passive operators (average)	Total operators (average)	Keyboard and mouse	Game controller	Gesture Recognition	Display screens
Team A	1	1.3	2.3	1			3
Team B	2	3	5	4			11
Team C	1	3	4	1			2
Team D	2	2.3	4.3	4			12
Team E	3.2	4	7.2	3			8
Team F	2.8	1.2	4	4			7
Team G	2.5	3.5	5.8	3	1		6
Team H	2.3	2.5	4.8	3		1	3

Table 10. Average number of operators, number of input devices, and number of display screens per team.

Table 10 shows that all teams had a heavy reliance on keyboards and mice for interacting with the robot system. Two of the eight teams also included alternative methods, Team G with a game controller and Team H with gesture recognition; however, these two teams were the lowest scoring teams in our study. While that result leaves us unable to suggest that the use of anything other than a keyboard and mouse would lead to the best performance, it is not necessarily the case. It is important to note that the interaction methods designed for the DRC Trials were used by the system developers, wherein the use of only a keyboard and mouse was most likely sufficient.

6.3 Control Methods

Each team used a variety of methods to control their robot on mobility and manipulation tasks, ranging from fine-grain individual joint control to more automated processes like waypoint placement. While almost all of the control methods were custom built as part of a team’s interaction method, there were common qualities that each shared, which we distilled into six categories of control methods:

- Individual joint control by typing numbers (JN): manually entering numbers to rotate individual joints
- Individual joint control using robot avatar (JA): rotating joints by manipulating a 3D robot avatar on screen by using a Cartesian rotation tool
- Pre-made script for action (PM): executing a pre-made script for a single action (e.g. sidestepping, crouching, etc.), with some variable input (such as distance, speed, etc.) depending on implementation
- Motion planning by 3D object linking (OBJ) (manipulation only): planning trajectories by linking a simulated limb of the robot to a 3D model that the operator places into the environment and then manipulating the model. When the model moves the simulated linked

robot limb moves with it and sends this trajectory to the robot

- Manual footstep placement (FP) (mobility only): placing individual footsteps on a surface
- Waypoint placement (WP): placing a waypoint or end goal for the robot to follow, with automated planning of footsteps/trajectories

Teams varied which control method they used by task and subtask. Use of one control method is not excluded from another; for example, Team D used pre-planned scripted actions to execute a task and adjusted movements during execution by manually placing footsteps.

An important difference between each of the control methods described above is the amount of input required by the operator to achieve a certain output. The amount of knowledge that the robot and/or operator must have about environmental features that could potentially affect performance (e.g., the location of the hurdle in the Terrain task, which could cause the robot to trip) is also important. We have defined three categories of interaction amount, defined by these factors. To more easily compare them we can look at an example wherein the output is held constant. For instance, consider a mobility task like commanding the robot to take 5 steps forward:

- Low amount of interaction: placing a single waypoint or end goal for the robot to walk to (WP). Footsteps and their accompanying trajectories are automated, and the robot plans around environmental features during/before execution as needed
- Medium amount of interaction: manually placing five footsteps for the robot to perform (FP). Trajectories for each footstep are automated, and the robot plans around environmental features during/before execution as needed
- High amount of interaction: using pre-made scripts to command the robot to step forward five times (PM). Trajectories for each footstep are automated based on input variables such as distance, speed, etc. from the operator, but the environmental features must be known to allow for proper planning, or the robot may have reactive features to adapt during execution as needed. Controlling individual joint angles either by manipulating a 3D simulation of the robot avatar or by typing numbers to achieve five footsteps (JN/JA), where nothing is automated, does require more interaction than PM, but for simplicity we are including it in this category. Every team in our study exhibited both JN/JA and PM, so the distinction between the two is unnecessary.

For a manipulation task, manual footstep placement (FP) for the medium amount of interaction category does not apply, and is instead replaced with motion planning by 3D object linking (OBJ). Trajectories for each arm and hand movement are automated, and the robot plans around environmental features during/before execution as needed. Also, the low amount of interaction category would still be waypoint/end goal placement (WP), but would take the form of placing a simulated robot hand pose or other goal identifier into the environment. For example, on the Valve task, Team F placed a shape over their acquired point cloud data to tell the robot where the valve wheel was, signifying the goal. The robot would then automate rotating the valve without any further operator input.

Note that this categorization does not take reliability into account, because that is specific to each team's implementation of the control method. Each team's performance dictates whether or not a control method was effective; one is not inherently better than the other. Table 11 shows each team's control methods for mobility and manipulation tasks, and the associated amounts of interaction.

	Control methods for mobility					Amount of interaction for mobility			Control methods for manipulation					Amount of interaction for manipulation		
	JN	JA	PM	FP	WP	High	Med	Low	JN	JA	PM	OBJ	WP	High	Med	Low
Team A	X	X	X	X	X	✓	✓	✓	X	X	X	X	X	✓	✓	✓
Team B	X	X	X	X	X	✓	✓	✓	X	X	X	X	X	✓	✓	✓
Team C	X	X	X	X	X	✓	✓	✓	X	X	X		X	✓		✓
Team D	X	X	X	X		✓	✓		X	X	X			✓		
Team E	X	X	X		X	✓		✓	X	X	X		X	✓		✓
Team F	X		X			✓			X	X	X		X	✓		✓
Team G	X	X	X	X		✓	✓		X	X	X			✓		
Team H	X		X			✓			X		X			✓		

Table 11. Mobility and manipulation control methods used by each team and their associated amounts of interaction.

7. Results and Discussion

Of the many analysis methods described above, we have found positive correlations between many of the metrics used to measure performance and to measure interaction features. Not all of our planned metrics ended up being used, but the pertinent and telling correlations are explained below. Most of the tables in this section have cells that are colored from lightest to darkest, which corresponds to the range of the data for each metric and is based on the magnitude of each individual data point within the dataset.

7.1 Critical Incidents

As described in Section 5.2, we coded critical incidents that occurred with the robots on the course during task runs. We divided these critical incidents into two categories. The first category clustered falls and resets, as these critical incident types led to a loss of 5 minutes from the team’s run time when an intervention was called or to the early end of a run, if less than 5 minutes remained when the intervention was called. The second category clustered the remainder of the critical incidents, none of which resulted in an immediate time penalty, although it certainly might have slowed the progress of the robot. All of the critical incidents for the observed teams and tasks are shown in Table 3 in Section 5.2.

	% of falls and/or resets	% of other critical incidents	% of all critical incidents
Team A	3.0%	2.3%	2.6%
Team B	6.1%	13.6%	10.4%
Team C	6.1%	15.9%	11.7%
Team D	15.2%	20.5%	18.2%
Team E	18.2%	9.1%	13.0%
Team F	21.2%	0.0%	9.1%
Team G	9.1%	13.6%	11.7%
Team H	21.2%	25.0%	23.4%

Table 12. Percentages of the aggregate of falls and/or resets vs. other critical incidents vs. all critical incidents per team. This table is sorted by the leftmost column, overall team performance.

In Table 12, we present a comparison of the number of critical incidents for each team compared to the total number of observed critical incidents, in falls and resets, other critical incidents, and overall. Teams who had a greater percentage of the falls and resets, incurring time penalties, fared worse overall than teams with a lower percentage of such critical incidents. The other critical incidents and overall critical incidents categories were less predictive of the team’s overall ranking.

7.2 Subtask Performance

As described in Section 5.3, we broke the DRC Trials Tasks into functional subtasks, resulting in a finer granularity than the scoring method used for each task. We then categorized each subtask as unobstructed traverse (UT), obstructed traverse (OT), first order manipulation (M) and second order manipulation (SOM; use of a tool).

Table 13 shows the percentages of successful attempts in these four categories for each team. This metric allows us to look at the success of teams on subtasks that they were trying to complete, rather than including all of the subtasks that were not attempted, either due to a lack of time or a lack of capability. We also show the percentages of successful attempts for both of the mobility subtasks (UT + OT) and both of the manipulation subtasks (M + SOM). Unsurprisingly, we see that, for the most part, the top teams had the higher percentages of success for both mobility and manipulation subtasks.

	Unobstructed traverse (UT)	Obstructed traverse (OT)	Mobility subtasks (UT and OT)	First order manipulation (M)	Second order manipulation (SOM)	Manipulation subtasks (M and SOM)
Team A	90%	100%	94%	100%	100%	100%
Team B	100%	0%*	73%**	100%	91%	94%
Team C	100%	71%	86%	100%	100%	100%
Team D	89%	67%	76%	100%	50%	80%
Team E	89%	38%	65%	100%	67%	89%
Team F	57%	82%	72%	33%	100%	50%
Team G	60%	67%	63%	100%	67%	80%
Team H	100%	0%	63%	0%	N/A	0%

Table 13. Average successful subtask attempts per team by team rank. As noted in Table 1, Team B was observed on Terrain, but we were not able to include the data in our analysis. If we had, * would be 50%, and ** would be 79%

The top four ranked teams in our study have the highest percentages for successful mobility subtask attempts (UT and OT) across all tasks. The top three ranked teams in our study also have the highest percentages for successful manipulation subtask attempts (M and SOM) across all tasks. To try to understand why Team E knocked Team D out of this ordering, we can test against a few additional axes. In Table 24 we can see that Team E required less interaction for manipulation than Team D. Given that Team E had fewer manipulation related falls and/or resets and other critical incidents than Team D, we can conclude that Team E’s control methods for manipulation were more reliable.

The values in Table 14 show a comparison between the average time required to complete the subtasks, separated by subtask function. The values show the percentage of time difference with respect to the aggregate. For example, the value for Team A in unobstructed traverse (UT) is -31%,

meaning that Team A used 31% less time for these subtasks than the average across all teams for these subtasks. A positive value in a cell means that the team performed that type of subtask more slowly than the average across all of that subtask type. This data was analyzed in an attempt to find correlations between situation awareness and task speed. We hypothesized that teams with greater situation awareness would tend to perform the tasks more quickly. However, no correlations could be made with the time breakdown shown here, suggesting that interaction methods and situation awareness did not impact speed. The DRC Trials did not reward speed with extra points, so there was no benefit to a team completing a subtask quickly, unless they had plans to complete additional subtasks in the overall task. This lack of reward for speed might have led to the lack of a correlation between speed and SA.

	Unobstructed traverse (UT)	Obstructed traverse (OT)	Mobility subtasks (UT and OT)	First order manipulation (M)	Second order manipulation (SOM)	Manipulation subtasks (M and SOM)
Team A	-31%	13%	-8%	1%	-9%	-5%
Team B	-16%	62%	26%	5%	12%	7%
Team C	35%	41%	33%	67%	140%	103%
Team D	-13%	-17%	-19%	-4%	30%	12%
Team E	5%	41%	22%	-20%	30%	5%
Team F	47%	-22%	2%	-37%	180%	72%
Team G	68%	196%	136%	99%	120%	108%
Team H	68%	N/A	N/A	N/A	N/A	N/A

Table 14. Percentages for teams that are above (positive) or below (negative) aggregate average time per subtasks. Larger negative values mean the team was much faster than average. Table sorted by team rank (left column).

The data in Table 15 shows differences in the points scored by the teams with respect to the aggregate, broken down by subtask. This analysis was performed to explore team capabilities for specific types of tasks and also to find any correlations between function capability and interaction methods. In general, team aptitudes for task type varied, though no correlations could be made to specific interaction methods. Of note is the correlation of the unobstructed traverse (UT) values to the team ranking; the faster walking teams for UT subtasks scored better in the overall competition. Four of the five teams who earned more than the average number of points for obstructed traverse (OT) were the top four ranked teams. First order manipulation has a similar finding for the top five ranked teams.

	Unobstructed traverse (UT)	Obstructed traverse (OT)	Mobility subtasks (UT and OT)	First order manipulation (M)	Second order manipulation (SOM)	Manipulation subtasks (M and SOM)
Team A	83%	104%	140%	92%	276%	167%
Team B	60%	31%	-20%	28%	41%	33%
Team C	37%	16%	-20%	60%	41%	52%
Team D	14%	31%	60%	28%	-53%	-5%
Team E	14%	2%	-20%	28%	-53%	-5%
Team F	-54%	-42%	-20%	-68%	-100%	-81%
Team G	-54%	-42%	-20%	-68%	-53%	-62%
Team H	-100%	-100%	-100%	-100%	-100%	-100%

Table 15. Percentages for teams that are above (positive) or below (negative) aggregate average points per subtasks. A large positive percentage indicates that the team was well above the average points scored for a subtask across all teams.

7.3 Utterance Coding

An analysis of the utterances between team members can provide us with some insight as to the success of or problems with their interaction design. We note that this analysis has some limitations. First, some teams spoke more than others; to account for this variability to the amount possible, we report results as a percentage of the total utterances made by the team over all observed tasks. Second, some of our observers recorded everything said in the control room, within the limitations of handwriting the notes, while other observers seem to have recorded less, judging from the number of utterances across different tasks for the same team when observed by different people. However, despite these limitations, we found some trends in the team utterances, which we report here. Such trends suggest that audio recording in the control room during runs, with later transcription, could be a valuable data set for analysis in the DRC Finals or other similar competitions.

7.3.1 Stress

Table 16 shows the percentage of a team’s overall utterances that were coded as being stressed. There are lower percentages of utterances coded as stressed for three of the top four teams (Teams A, B, and D). However, the other top team, Team C, had the highest percentage of stressed utterances, a majority of which were recorded during the Door task. We believe that this anomaly was due to the fact that the team used a different operator for this task than for all of the other tasks, a fact explicitly told to the observer in the control room before the run started.

The amount of stress in a team might be caused by the team knowing that the run is going badly, or might be a predictor of a run about to go badly. One possibility for learning about the causality between the two would be to use biometrics, which could allow for timing comparisons of when stress rose in the operators and when the robot had critical incidents. If stress rises before the critical incident occurs, it could be attributable to the interface design. More importantly, if stress was found to occur before the critical incident, it might be avoided through better interface design. Such investigations are an open problem for human-robot interaction.

	% of total utterances coded as stressed
Team A	6.0%
Team B	5.2%
Team C	15.2%
Team D	4.0%
Team E	8.5%
Team F	6.2%
Team G	11.4%
Team H	12.9%

Table 16. Percentage of a team’s overall utterances that were coded as being stressed.

7.3.2 Comms

On the Terrain task, six of the eight teams scored points. The teams who scored points had no utterances about comms at all. Teams F and H scored no points and had 12.2% and 5.5% of their total utterances on the Terrain task about comms, respectively. A similar pattern is revealed for the Wall task, where the two teams who scored points (Teams A and C) had no utterances about comms, while Teams D, E, and H had 2.2%, 20.7%, and 10.0% of their total utterances on the Wall task about comms, respectively. The pattern almost holds for Door, where we observed all of the eight teams; three of the four teams who scored points had no utterances about comms. This data can be seen in Table 17.

	Points scored on Terrain	% of total utterances about comms during the Terrain task		Points scored on Wall	% of total utterances about comms during the Wall task		Points scored on Door	% of total utterances about comms during the Door task
Team A	4	0.0%	Team A	4	0.0%	Team A	4	0.0%
Team B	3*	0.0%*	Team C	1	0.0%	Team B	1	0.0%
Team D	2	0.0%	Team D	0	2.2%	Team C	1	1.0%
Team C	1	0.0%	Team E	0	20.7%	Team G	1	0.0%
Team E	1	0.0%	Team H	0	10%	Team D	0	0.0%
Team G	1	0.0%				Team E	0	6.3%
Team F	0	12.2%				Team F	0	2.3%
Team H	0	5.5%				Team H	0	6.6%

Table 17. Teams in order of points scored for the Terrain (Left), Wall (Center), and Door (Right) tasks vs. the percentage of total utterances about comms during each task, respectively. *Observed values shown in this table, although excluded from other analyses, for a reason withheld to preserve anonymity.

These results indicate that teams performed better when they did not have to spend time talking about the variable data rate and delay. Team A designed their system to only require the amount of bandwidth available at the lower level, and thus they never had to be concerned with the comms level. Some teams, such as Team C, reduced their camera feeds to grayscale to limit the amount of bandwidth needed. Looking at the overall utterances for teams across all of the tasks on which they

were observed, three of the top four teams (Teams A, C, and D) had fewer utterances about comms and also had no utterances about comms that were coded as negative, as seen in Table 18.

	% of total utterances about comms	% of total negative utterances about comms
Team A	0.0%	0.0%
Team B	4.6%	0.0%
Team C	0.8%	5.9%
Team D	3.4%	0.0%
Team E	6.3%	0.0%
Team F	9.3%	5.3%
Team G	4.6%	0.0%
Team H	6.2%	18.2%

Table 18. Percentage of total utterances about comms vs. negative utterances about comms, sorted by team performance.

7.3.3 Talking about task related items instead of robot control

The ability of a team to discuss aspects of the task at hand, rather than focusing on the appropriate methods for controlling a robot, should lead to better performance. We found that teams who had more utterances coded as obstacle, which meant the walls and other unmovable parts of the course, or tool, which meant the pieces of wood acting as debris, the handles of the valves, and the drill, had better performance on the task.

On the Debris task, two of the four observed teams, A and B, scored points; Team A had 28.6% and Team B 18.2% of their utterances coded as obstacle, while teams not scoring points had no utterances in this category, as seen in Table 19. We saw a similar split for tool utterances, with Team A having 28.6% and Team B 45.5% of utterances coded as tool, but 0% and 2.7% for non-point scoring teams.

On the Wall task, two of the five observed teams scored points (A and C); teams with points had 19.4% and 22.2% of their utterances coded in the obstacle category (the wall itself and table that held the drill were considered obstacles in our coding) and the three teams who did not score points had 0%, 0%, and 5.6% (see Table 19).

	Points scored on Debris	% of total utterances about the obstacle during the Debris task	% of total utterances about the tool during the Debris task
Team A	1	28.6%	28.6%
Team B	1	18.2%	45.5%
Team F	0	0.0%	0.0%
Team G	0	0.0%	2.7%

	Points scored on Wall	% of total utterances about obstacle during the Wall task
Team A	4	19.4%
Team C	1	22.2%
Team D	0	5.6%
Team E	0	0.0%
Team H	0	0.0%

Table 19. Left: Teams observed on the Debris task, in order of points, with the percentage of total utterances coded as about an obstacle and percentage of utterances coded as about a tool. Right: Teams observed on the Wall task, in order of points, with the percentage of total utterances coded as about an obstacle.

For tasks that involve tools (Debris, Door, Hose, Valve, and Wall), the top four teams had the highest percentages of total utterances about the tool (e.g., the piece of wood in Debris, the drill in Wall, the valve handle and wheel in Valve), which is likely to be an indication that the team was able to focus on the task at hand rather than robot or interface mechanics.

We expect a correlation between this result and the use of manipulation control methods that involve the operator directly interacting with the tool, such as waypoint placement (wherein the operator places a waypoint onto the tool) and motion planning by 3D object linking (wherein the operator manipulates a 3D model of the tool), and the use of simulated 3D models of objects within a data display. Table 20 shows the inclusion of these additional axes, which shows the correlation. Note that Team E is within the top four in this ordering by percentage of utterances about tool. Table 20 provides evidence as to why Team E knocked out Team D for this ordering: Team E used simulated 3D models of objects (many of which are the subject of tool utterances) as part of their interaction method, whereas Team D did not, involving much less direct interaction/consideration of the tools being used. If we also look at the percentage of manipulation related falls and/or resets, we see that Team D had 3.0% to Team E's 0%, which removed time that could have been used to score points.

	% of total utterances about the tool on relevant tasks	Sim objects or obstacles	Waypoint placement	Motion planning by 3D object linking	Team percentage of the aggregate of manipulation related falls and/or reset
Team B	26.5%	X	X	X	0.0%
Team E	12.7%	X	X		0.0%
Team A	12.3%	X	X	X	0.0%
Team C	11.9%		X		0.0%
Team D	6.9%				3.0%
Team H	6.5%				6.1%
Team F	5.1%	X	X		6.1%
Team G	1.3%				0.0%

Table 20. Percentage of total utterances about the tool on relevant tasks vs. tool interaction data displays and manipulation control methods vs. manipulation related falls and/or resets. Teams are ordered according to their percentage of total utterances about the tool on relevant tasks.

7.4 Sensor Fusion

Teams can be compared by the amount of sensor fusion their interface utilized in presenting the data back to the operator. Based on the number of data displays that were combined to form unique instances of sensor fusion they can be split into three categories: high (Team A), medium (Teams B through E), and low (Teams F through H). The top four teams in our study all utilized high or medium sensor fusion, while the bottom four teams all utilized medium or low sensor fusion.

If we compare all teams across their amount of sensor fusion and their average successful subtask attempts, shown in Table 21, we see the top four ranked teams had the highest average successful subtask attempts for mobility. For manipulation, the top three ranked teams had the highest averages; Team D got knocked out by Team E. Both teams had the same level of sensor fusion (medium), but if we also look the amount of falls and/or resets, as seen in Table 21, Team E had more falls and/or resets than Team D, suggesting why Team E was not ranked in the top four.

	Amount of sensor fusion	Average successful mobility subtask attempts (UT and OT)	Average successful manipulation subtask attempts (M and SOM)	Team percentage of the aggregate for falls and/or resets
Team A	High	94%	100%	2.9%
Team B	Medium	73%	94%	11.4%
Team C	Medium	86%	100%	5.7%
Team D	Medium	76%	80%	14.3%
Team E	Medium	65%	89%	17.1%
Team F	Low	72%	50%	20.0%
Team G	Low	63%	80%	8.6%
Team H	Low	63%	0%	20.0%

Table 21. Amount of sensor fusion vs. successful subtask attempts for mobility and manipulation vs. falls and/or resets.

Three teams (A, B, and E) used foveated vision as part of their interface. Those three teams also used simulated 3D models of obstacles and objects in the environment to plan robot movements. Given that foveated vision gives the operator a wider view of the robot’s environment it is expected that less critical incidents would occur. Also, the use of simulated 3D models of objects should provide more accurate planning, given that the characteristics of the model are known to the system rather than having to rely solely on point cloud data or camera views, assuming the models are placed into the environment correctly, thus also resulting in less critical incidents. Team F also used simulated 3D models, but given their low percentage of successful manipulation subtask attempts, we can assume that other factors contributed (Team F was observed falling twice on the Valve task due to what appeared to be balance issues after manipulating the valve wheel). Compared to the other teams, teams A, B, and E had a statistically significant faster subtask time. On average, these teams completed the subtasks in 75% of the time taken by the other teams ($p < 0.01$). If we also include additional axes for falls and/or resets and other critical incidents, as seen in Table 22, we can see that Teams A, B, and E had no manipulation related falls and/or resets, and had the least amount of other critical incidents after Team F. However, Team F had the highest falls and/or resets, which are more detrimental to performance than other critical incidents.

	Foveated vision	Sim objects or obstacles	Team percentage of the aggregate of manipulation related falls and/or resets	Team percentage of the aggregate of manipulation related other critical incidents	Average time to complete manipulation subtasks
Team A	X	X	0.0%	2.3%	-5%
Team B	X	X	0.0%	11.4%	7%
Team E	X	X	0.0%	2.3%	5%
Team C			0.0%	16.0%	103%
Team D			3.0%	20.5%	12%
Team F		X	6.1%	0.0%	72%
Team G			0.0%	13.6%	108%
Team H			6.1%	25.0%	N/A

Table 22. Foveated vision and simulated 3D models of objects vs. manipulation related falls and/or resets and other critical incidents vs. average time to complete manipulation subtask above or below the aggregate.

7.5 Control Methods and Interaction Amount

Lower interaction amounts require less effort from the operator and reduces his/her cognitive load by abstracting the level of detail to which commands must be entered. With lower cognitive load, we would expect higher performance. There is a correlation between average successful subtask attempts and lower interaction amounts, for both mobility and manipulation, respectively, as seen in Tables 23 and 24. Given that each team created their own control methods that fall under the medium and low categories interaction amount (some high interaction amount control methods were not custom made and came with the robot platform), we can infer whose were more effective than others (such as the ordering of Teams A, B, C, and E for mobility).

However, there are outliers in each. While the top four ranked teams are also had the highest success rates for mobility subtasks, Team E used the same interaction amount as Teams A, B, and C, but is in the bottom half for success rate. We can further investigate this issue by including an additional axis of average total operators (active and passive) and mobility related critical incidents, seen in Table 23. With the addition of these variables, we can attribute Team E’s poor mobility performance to having the highest number of average operators. This high number of operators could have resulted in a “too many cooks in the kitchen” scenario; while more operators can spread out cognitive load, too many can make the information difficult to manage. We can also see that Team D had less falls and/or resets and other critical incidents than Team E, which could be attributed to the number of operators and/or differences in control method implementation. While Team E exhibited low interaction amounts, their custom made mobility control methods were most likely less accurate and reliable than Team D’s which required medium interaction amounts.

The top three ranked teams also had the highest success rate for manipulation subtasks, with Team E knocking Team D out of its scoring rank. This finding makes sense, given that Team D’s interaction amount is higher than Team E’s. By additionally comparing against manipulation related critical incidents, as seen in Table 24, we can see that Team D had more falls and/or resets and other critical incidents than Team E, the inverse of their performances in mobility. The difference in interaction amounts supports this finding. It is important to note that our coverage of tasks per team is not a factor when comparing Team D and E, as we observed the same tasks for both teams.

	Average successful mobility subtask attempts	Amount of interaction for mobility			Percentage of mobility related falls and/or resets	Percentage of mobility related other critical incidents	Average total number of operators
		High	Med	Low			
Team A	94%	✓	✓	✓	3.0%	0.0%	2.3
Team C	86%	✓	✓	✓	6.1%	0.0%	4
Team D	76%	✓	✓		12.1%	0.0%	4.3
Team B	73%	✓	✓	✓	0.0%	2.3%	5
Team F	72%	✓			15.2%	0.0%	4
Team E	65%	✓		✓	18.2%	6.8%	7.2
Team G	63%	✓	✓		9.1%	0.0%	4.8
Team H	63%	✓			15.2%	0.0%	5.8

Table 23. Average successful mobility subtask attempts per team vs. amount of interaction for mobility vs. mobility related falls and/or resets and other critical incidents vs. average total number of operators. Teams are ordered according to their average successful mobility subtask attempts.

	Average successful manipulation subtask attempts	Amount of interaction for manipulation			Percentage of manipulation related falls and/or resets	Percentage of manipulation related other critical incidents	Average total number of operators
		High	Med	Low			
Team A	100%	✓	✓	✓	0.0%	2.3%	2.3
Team C	100%	✓		✓	0.0%	16.0%	4
Team B	94%	✓	✓	✓	0.0%	11.4%	5
Team E	89%	✓		✓	0.0%	2.3%	7.2
Team D	80%	✓			3.0%	20.5%	4.3
Team G	80%	✓			0.0%	13.6%	4.8
Team F	50%	✓		✓	6.1%	0.0%	4
Team H	0%	✓			6.1%	25.0%	5.8

Table 24. Average successful manipulation subtask attempts per team vs. amount of interaction for manipulation vs. manipulation related falls and/or resets and other critical incidents vs. average total number of operators. Teams are ordered according to their average successful manipulation subtask attempts.

8. Lessons Learned

The DRC Trials provided the first opportunity to study strategies for the design of human-robot interaction for legged robots, most of them humanoid, that were required to complete disaster response tasks. Before this event, there had never been such a collection of humanoid or legged robots in a single location to complete the same set of tasks.

At a high level, we found that teams needed to be capable in the following areas to be successful:

1. Robot mobility,
2. Robot manipulation,
3. Situation awareness of the robot and its surroundings, and
4. An effective way to command the robot.

Robot mobility and manipulation both were improved when control methods with lower interaction amounts were used (see Tables 23 and 24). Having the ability to command a robot to a waypoint or being able to show the robot where to place its feet are much more automated processes than individual joint control; these higher levels of control require much less effort from the operator, and, when implemented correctly, can make the HRI more efficient and less prone to errors. For example, Team F moved its robot on the Terrain task by typing numbers in text boxes to change joint angles. When the operator wanted to move between static and dynamic walking, the operator needed to press a button on the interface. When the operator forgot to press the button, the robot fell. A lower cognitive load on the operator might have prevented this error. Instead, much of the effort applied by the team was in controlling the robot’s mobility instead of focusing on the task at hand.

However, as we see in the data, it is not enough for the robot’s control methods to enable it to be highly capable in the areas of mobility and manipulation. Teams A, B, C and E all exhibited low interaction amounts for mobility, and Teams A, B, C, E, and F all exhibited low interaction amounts for manipulation. Yet Team A ranked higher than all other teams. The difference comes in the interface design.

Team A had the highest amount of sensor fusion. In fact, they were able to display everything needed to control the robot on a single monitor. They had a second monitor duplicating the fused information, which only used by the supervisor. A third monitor, to the left of the operator's primary monitor, was only used to start up and monitor the robot's controllers as needed. This method enabled them to efficiently maintain SA, noted by their lack of critical incidents (see Table 14) and use of a single active operator.

In contrast, many teams had multiple operators with each operator interacting with multiple data streams, either fused or unfused. A lack of sensor fusion and lack of operator fusion is detrimental to the overall performance (see Table 26). For a single operator with multiple data streams, the operator must interpret all of the separate streams to build situation awareness – the essence of decentralized SA. This problem is further compounded when multiple operators have different information that must be fused in order to effectively control the robot. For example, Team E's instance of sensor fusion was propagated throughout many monitors in different forms. The left and right operators were able to adjust the level of sensor fusion on their displays as needed, but the main operator viewed the available data spread out across three monitors, splitting the camera feed, 3D point cloud with robot avatar, and 2D local distance visualization. This interface produced a lack of operator fusion, causing their performance to degrade.

We saw many instances in the utterances by teams where one operator would correct another operator about a direction in an utterance (e.g., "Move left and forward." "You mean right and forward?" "Yes."). The lack of sensor fusion as well as the lack of operator fusion leads to diminished SA, leading to more critical incidents and lower performance. The average number of operators was predictive of the overall rankings of our study participants, with the top half of teams having 1 or 2 active operators while the bottom half had an average of more than 2 active operators, with a high of 3.2 for Team E (see Table 10).

All eight teams used some number of keyboards and mice. The two bottom ranked teams added alternative interaction methods in addition to their keyboards and mice (Team G had a game controller to place robot footsteps and Team H had a gesture control device for arm movements). While this result leaves us unable to suggest that the use of anything other than a keyboard and mouse would lead to the best performance in this case, this is not necessarily the case. It is important to note that the system developers used the interaction methods designed for the DRC Trials; keyboards and mice can be sufficient for developer interfaces. It is also worth noting that development time for some teams was very short, resulting in incomplete interfaces.

Human-robot interaction designed for the first responders who one day would be the users of these systems would need be significantly different. Even for the top ranked teams, it would take an impractical amount of training before a non-roboticist first responder would be able to control the robot effectively, let alone proficiently. As these types of robot systems mature from development to use, we expect to see changes in the interaction methods and suggest that developers start to create these interaction methods now, rather than after the robot hardware and controllers are fully mature. Integrating the design and development of the human-robot interaction with the design and development of the robot system will lead to higher performing, easier to use systems.

Yanco, Drury and Scholtz [2004] also discussed the fact that HRI evaluations conducted in a robotics competition setting have traditionally involved robot operation by developers, rather than end users. However, their study also included a domain expert, a special operations fire chief

trained in search and rescue and with experience operating robots. A comparison of task performance across the expert user and the developers participating in formal competition trials revealed that the developers relied more heavily on sensor data while the expert was more dependent on live video feed data. While the current DRC Trials study involved only developers as operators, this finding is relevant to the interpretation of our results in that the primary sources of information used may not translate to end users. Within the current study, operators often relied on point cloud, simulation data, and sensor data independently or in conjunction with video data. Furthermore, Yanco, Drury and Scholtz [2004] found that more sensor types did not necessarily increase awareness, especially if the sensor data was not well fused into information for the operator. The success or failure of multiple sources of sensor data by DRC teams may have been mediated by the quality of data fusion and presentation.

With the move from wheeled or tracked robots to walking robots, mobility and balance is now as much of an issue for robot control as manipulation has been over the past decades. For example, Team F on the Valve task appeared to be able to manipulate the large valve wheel with ease, using both arms and hands simultaneously, but upon releasing their grip the robot fell forward. At this stage of robot development for a tracked or wheeled robot, an operator generally moves a joystick in a direction that makes the robot drive in that direction, with few worries about stability when not on extreme terrain. However, legged, especially bipedal, mobility is much more complicated, with many joints moving in tandem to produce a single step in a direction, generally moving very slowly, for which balance must always be accounted. An operator cannot control all of these variables individually, but rather must rely on automated kinematic control to produce a proper trajectory to move the robot. The lowest interaction amount for mobility, waypoint placement, is a good example of removing the need for the operator to worry about individual joint angles, footstep placement, and keeping the robot balanced. To reach the same level of proficiency currently attained in a tracked robot platform, allowing the operator to think about the movement of the bipedal robot in more directional manner, a much lower amount of interaction and autonomy than what is currently available, is needed, based on our observations of the participating teams.

Our analysis showed that, of the tasks we observed, the subtasks whose functions were categorized as “obstructed traverse” were, on average, more difficult than the other types of subtasks. The Terrain task consists entirely of obstructed traverse subtasks (save for walking to the finish line, if applicable), and had the highest number of falls (see Table 3). Even more telling is that, for the teams we observed, the percentage of the successfully completed mobility subtasks (unobstructed traverse and obstructed traverse combined together) predicted if a team would be in the top half or bottom half (see Table 15). These humanoid/legged robots did make a step forward in mobility with the ability to navigate a steep stairwell, a task not easily accomplished with today’s existing systems. However, there was a technological step backward made with the ability to cross hurdles, a task easily accomplished with numerous existing robotic platforms. In these early years of the technology, we will need to balance the leaps in technical capability of complex operations (such as heavy debris removal and ascending steep stairwells) with similar advances in existing, straightforward tasks (such as traversal over minimal obstructions).

The more successful teams used mobility control methods that required low or medium amounts of interaction in which the operator could place a point or points in the environment, to which the robot would then walk. The success of these methods is mostly due to the fact that the task environments were static. If the environments were dynamic, say with malleable terrain such as mud or with blocks that could be in different locations, the mobility control methods would likely

not have performed as well, due to a lack of autonomous capabilities during execution. We observed that one team had completely planned its route through the terrain in advance by having their team members measure distances between each obstacle on the Terrain task and make the plan. Such access would not be possible in during a real-world disaster response scenario.

With the addition of degraded communications, some amount of autonomy is even more desirable. In the face of degraded communications, the most successful team (Team A) always used a data rate low enough to communicate to their robot at the same speed during good and bad comms, which allowed them to have their peak performance during the entire run, rather than every other minute. As the DARPA Robotics Challenge moves towards the finals, where the tasks will run end to end and the competition will include a surprise task, the robots' autonomous capabilities will need to increase.

If we take the DRC Trials as an indicator of the future of USAR robotic platforms, we are seeing HRI move from the ratio of one operator to one or more robots, often desired by members of the military, to multiple operators to a single robot. Recognizing that these systems are very complex, what could be done to reduce the number of operators needed? One possibility would be to reduce the amount of HRI needed for mobility. Teams A and C had one active operator, so it is possible for a bipedal robot to not only be put through the DRC Trials tasks with a single operator, but also for it to be done with a top ranked score. Team B had two operators: one was building models and the other planning the robot's motions. Likely, these two positions could be combined into one as well. The fact that the tasks were being run in a competition was probably a contributing factor to the increased number of operators. Prior competitions have included the number of operators in denominator of the scoring metrics (e.g., AAI/RoboCup Rescue in 2001 and 2002). If the same were done for the DRC, would we see a great improvement in performance? Would this scoring change, penalizing teams for larger number of operators, drive teams to streamline their interfaces?

8.1 Categories of Operator Effort

Throughout the observations and data analysis, we identified a common theme associated with the human-robot interaction techniques and performance, which was focused on how effort was applied towards completing the task. We explored this theme in an attempt to generalize our findings, but, although it is derived from the data, it is unable to be directly supported. We have included this discussion because we believe it is effective at uniting and providing greater context of the data and analysis described previously.

We found that operator effort applied to completing a task fell into three categories. These categories range from the lower-level robot control and developing situation awareness to the highest level of interacting with the environment. These categories build on each other such that effective lower-level control is required for effective higher level interaction. Our evaluation is in agreement with this as we found that teams that were able to apply more operator effort at the higher levels performed better than teams that only applied their operator effort at the lower levels. To be effective at completing a task, the team needs to apply their effort at this highest level as opposed to the lower levels of control and SA. Operator effort is applied to one of these categories:

1. Interacting with the robot through the interface,
2. Developing an understanding of the environment and how the robot can interact with it, and
3. Interacting with the environment.

To be effective at the first category, the teams required centralized SA and streamlined methods for controlling the robot. Robot automation reduces the amount of effort required in this category. Additionally, methods of sensor fusion helped the teams to understand the robot and its place within the environment. Some teams attempted to compensate for a lack of effective SA by increasing the number of operators, but we found that this was ineffective because of the lack of centralized SA. Bandwidth issues also fall within this discussion, as it provides a distraction for operators when they should be focusing on the task at hand.

The second category consists of enabling the robot to interact with the environment. This requires interpreting the environment in which it is operating, locating things like debris and valves and determining how best to interact with them. The top teams were the ones that were effective at generating this understanding such that they could effectively command the robot to interact with it. These teams used tools such as combining live video with a simulation to provide this understanding. By placing virtual objects with the properties of the objects in the real world (e.g., the dimensions and dynamics of a valve wheel), teams conveyed this understanding to the robot. Less effective teams applied a much greater effort here and subsequently had less effort to apply to actually interacting with the environment.

The third category is operator effort applied to completing the task at hand. The most effective teams are the ones that were able to apply their operator effort in this category as it directly converts operator effort towards task completion. Effective implementations of control methods that exhibited medium and low amounts of interaction for mobility and manipulation (such as those methods discussed in section 6.3) saw those teams applying more effort in this category. For successful teams, little operator effort is required for the first two categories, enabling more effort to be applied in this category. In the circumstance in which all effort is applied at this level, the robot has become an extension of the human him/herself, with no constraints in the flow of information through the interface. The interface here is effectively invisible in terms of robot control.

8.2 Design Guidelines

We propose the following design guidelines for human-robot interaction with humanoid/legged robots for disaster response and other situations with remote interaction, noting that there is some overlap with the HRI design guidelines suggested by evaluations conducted of USAR studies with tracked and wheeled robots (e.g., [Yanco, Drury & Scholtz 2004; Yanco and Drury 2007]).

Increase sensor fusion. Build an effectively fused representation of the data streams rather than require an operator to do the same thing mentally using separate windows. Better sensor fusion reduces the cognitive load for the operator, freeing him or her up for other tasks.

Decrease the number of operators. The use of more operators leads to the need for more collaboration, which opens the door to increased misunderstandings and confusion. Centralized SA is a must for effective operation.

Decrease the amount of operator input needed to control the robot. More automated capabilities of the robot can improve performance by reducing the amount of effort required to address the lower-level tiers described above. If the robots have greater capabilities for mobility and manipulation, the human-robot interaction can occur at a higher level and effort can be applied to

the higher tiers. We have seen that lower amounts of interaction, when implemented effectively, led to better performance. However, most of the interaction amounts exhibited are not very far from teleoperation. With even more automation, the HRI can be designed in a way that will be faster (e.g., showing where the robot should end up rather than placing every single step in the interface), less prone to errors (e.g., mistyping a number in a text box for joint control), and enable the operator to focus on the higher-level tiers, building an understanding of the environment and interacting with it.

Don't separate the robot into legs and arms. Most teams had a mobility phase and a manipulation phase, switching back and forth as necessary. If we stick to this division, we will fail to create a unified robot system. A division of task types create breaks in the SA and keeps operators narrowly focused rather than maintaining higher-level awareness.

Plan for low bandwidth. In the DRC Trials, teams were given advance notice of the bandwidth restrictions and the manner in which it would be implemented. The highest ranked teams either designed their system to work at the low bandwidth all of the time or had a method for only pulling high-resolution data during the periods of high bandwidth.

Design for the intended users. While fully understanding the complexity of the robot systems, human-robot interaction that does not consider the end user at the start of the development of the robot system will never be as effective as HRI that is designed as such from the beginning. It's not just the robot capabilities that shape the HRI, but the capabilities and needs of the human that should shape the HRI and the robot capabilities as well.

9. Planned Evaluation Improvements for the DRC Finals

During the evaluation of the DRC Trials, we have noted some areas where improvements could be made in future evaluations. We expect to complete a similar evaluation for the DRC Finals, and intend to incorporate as many of these evaluation improvements as possible.

Improved understanding of each team's interface. This knowledge would allow us to understand what each team was doing at a more detailed level and how they were interpreting their feedback. Additionally, it might allow us to comment with more detail on missed critical information and situation awareness during task execution. Such understanding could be achieved through visits with teams prior to the competition, given that teams are unlikely to have time once on site.

More detailed tracking of control methods per task. Having the ability to track each specific control method used by a team per task would make our analysis much more impactful, as we could better call out techniques for each subtask type. A better understanding of each team's interface will allow these measures to be more easily tracked during the competition.

Increased detail on team utterances. We will focus on methods to categorize operator interaction, and develop our note-taking sheets to assist in this process. Electronic recording of the dialog would assist as well (assuming teams would agree to it).

Track the data bandwidth changes more closely. In the DRC Finals, DARPA intends to continue the inclusion of a variable bandwidth choke on the system. Tracking these changes automatically during observations would help determine if issues arose as a result of these bandwidth fluctuations.

Ultimately, evaluation of robot competitions is limited by the need to allow teams to compete without interference. However, such competitions provide a unique opportunity to directly compare a relatively large number of systems designed for the same task. The findings of such studies can lead to the development of improved robot systems, allowing the community to learn the lessons of many teams at once and in comparison to one another, rather than reports of individual results.

10. Acknowledgements

In memory of Mike Stilman and Seth Teller.

This research has been supported in part by DARPA under W31P4Q-13-C-0136 and W31P4Q-13-C-0196. The views expressed are those of the authors and do not reflect the official policy or position of the Department of Defense or the U.S. Government. Approved for Public Release, Distribution Unlimited.

Many people have contributed to the research described in this paper. Gill Pratt and Eric Krotkov at DARPA provided much assistance in arranging our study of the DRC Trials. Wendell Sykes and Betsy Constantine participated in the study planning and were part of our observation team at the DRC Trials. Ann Virts and Adam Jacoff, both of the National Institute of Standards and Technology (NIST), provided support to our evaluation team, both before and during the event. John Blich and Cori Lathan assisted with study planning on the practice day of the DRC Trials. Lisa Barniecki assisted with data analysis. Finally, many thanks to all of the teams who participated by providing information, answering questions, and allowing us to observe during the competition.

References

[Adams 2002] J. A. Adams. “Critical considerations for human-robot interface development.” *Technical Report of the AAAI Fall Symposium on Human Robot Interaction*, pp. 1-8, 2002.

[BDI 2014] Boston Dynamics. “Boston Dynamics: Atlas - The Agile Anthropomorphic Robot,” http://bostondynamics.com/robot_Atlas.html, accessed February 2014.

[DARPA 2014] DARPA. “DRC Trials Score Analysis, Anonymous”, <http://www.theroboticschallenge.org/dashboard/scoreboard/DRCTrialsScoresAnalysisAnonymousv11DISTAR22423.pdf>, accessed March 2014.

[DRC 2014]. DARPA. “DARPA Robotics Challenge Website,” <http://darparoboticschallenge.com/>, accessed February 2014.

[DRC Videos 2014] “DARPA YouTube Channel.” <http://www.youtube.com/user/DARPAtv>, accessed February 2014.

[Diffler et al. 2003] M.A. Diffler, E.L. Huber, C.J. Culbert, R.O. Ambrose, and W.J. Bluethmann. “Human-robot control strategies for the NASA/DARPARobonaut,” *Proceedings of the 2003 IEEE Aerospace Conference*, Vol. 8, pp. 8_3939 – 8_3947, 2003.

[Fong, Thorpe, and Baur 2002] T. Fong, C. Thorpe, and C. Baur. “Robot as partner: vehicle teleoperation with collaborative control.” In Multi-Robot Systems: From Swarms to Intelligent

Automata, Springer, pp. 195-202.

[Fong et al. 2013] T. Fong, J.R. Zumbado, N. Currie, A. Mishkin, and D.L. Akin, "Space telerobotics: unique challenges to human-robot collaboration in space," *Reviews of Human Factors and Ergonomics*, Vol.9, pp. 6-56, 2013.

[Glaser and Strauss 1967] B. Glaser and A. Strauss. *The Discovery of Grounded Theory: Strategies for Qualitative Research*. Chicago: Aldine Pub. Co., 1967.

[Nourbakhsh et al. 2005] I. R. Nourbakhsh, K. Sycara, M. Koes, M. Yong, M. Lewis, and S. Burion. "Human-robot teaming for search and rescue." *IEEE Pervasive Computing*, 4(1): 72-78, 2005.

[Micire 2010] M. J. Micire, "Multi-Touch Interaction for Robot Command and Control," PhD Thesis, University of Massachusetts Lowell, Lowell, MA. December 2010.

[Murphy 2014] R. R. Murphy. Disaster Robotics. Cambridge MA: The MIT Press, Intelligent Robotics and Autonomous Agents series), 2014.

[Murphy and Burke 2005] R. R. Murphy and J. L. Burke, "Up from the rubble: lessons learned about HRI from search and rescue." *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 49(3): 437-441.

[Scholtz 2002] J. Scholtz. "Evaluation methods for human-system performance of intelligent systems." *Proc. of the Performance Metrics for Intelligent Systems (PerMIS) Workshop*, 2002

[Scholtz et al. 2004] J. Scholtz, J. Young, J. L. Drury, and H. A. Yanco. "Evaluation of human-robot interaction awareness in search and rescue." *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, Vol. 3, pp. 2327-2332, April 2004.

[Sian et al. 2002] N.E. Sian, K. Yokoi, S. Kajita, F. Kanehiro, and K. Tanie, "Whole Body Teleoperation of a Humanoid Robot - Development of a Simple Master Device using Joysticks." *Proceedings of the 2002 IEEE/IRSI Intl. Conference on Intelligent Robots and Systems* Lausanne, Switzerland, October 2002, pp. 2569 – 2574.

[Yanco and Drury 2007] H. A. Yanco and J. L. Drury. "Rescuing interfaces: a multi-year study of human-robot interaction at the AAI Robot Rescue Competition." *Autonomous Robots*, 22(4): 333-352, 2007.

[Yanco, Drury & Scholtz 2004] H. A. Yanco, J. L. Drury, and J. Scholtz. "Beyond usability evaluation: analysis of human-robot interaction at a major robotics competition." *Journal of Human-Computer Interaction*, 19(1&2): 117-149, 2004.