Creating Metrics for Human-Agent Teams

Elizabeth Cavanah
Iowa State University, ecavanah@iastate.edu

Zachery Ford
Iowa State University

Angelica Jasper
Iowa State University, amjasper@iastate.edu

Jacklin Stonewall
Iowa State University, jacklins@iastate.edu

Stephen B. Gilbert
Iowa State University, gilbert@iastate.edu

See next page for additional authors

Follow this and additional works at: https://lib.dr.iastate.edu/imse_conf

Part of the Ergonomics Commons, and the Operational Research Commons

Recommended Citation
https://lib.dr.iastate.edu/imse_conf/219

This Conference Proceeding is brought to you for free and open access by the Industrial and Manufacturing Systems Engineering at Iowa State University Digital Repository. It has been accepted for inclusion in Industrial and Manufacturing Systems Engineering Conference Proceedings and Posters by an authorized administrator of Iowa State University Digital Repository. For more information, please contact digirep@iastate.edu.
Creating Metrics for Human-Agent Teams

Abstract
This paper describes the requirements and overall design of a testbed for a human-agent team. A robust, flexible testbed will enable researchers to evaluate the effectiveness of human-agent teaming concepts and issues. In the testbed, MazeWorld, multiple agents play different roles in which tasks are interdependent. Each role can be served by a human or an autonomous agent. Metrics were developed to capture individual and team effectiveness and allow researchers to compare different types of teams and teamwork protocols. Examples of the current task design are presented with the discussion of the future development of the system.

Disciplines
Ergonomics | Operational Research

Comments

Authors
Elizabeth Cavanah, Zachary Ford, Angelica Jasper, Jacklin Stonewall, Stephen B. Gilbert, and Michael C. Dorneich

This conference proceeding is available at Iowa State University Digital Repository: https://lib.dr.iastate.edu/imse_conf/219
Creating Metrics for Human-Agent Teams

Elizabeth Cavanah
Zachery Ford
Angelica Jasper
Jacklin Stonewall

Stephen Gilbert, Ph.D
Michael Dorneich, Ph.D

Iowa State University

ABSTRACT

This paper describes the requirements and overall design of a testbed for a human-agent team. A robust, flexible testbed will enable researchers to evaluate the effectiveness of human-agent teaming concepts and issues. In the testbed, MazeWorld, multiple agents play different roles in which tasks are interdependent. Each role can be served by a human or an autonomous agent. Metrics were developed to capture individual and team effectiveness and allow researchers to compare different types of teams and teamwork protocols. Examples of the current task design are presented with the discussion of the future development of the system.

INTRODUCTION

As autonomous vehicles and software agents have become more integrated into society, the study of human-agent teams (HATs) has become more necessary. Combining agents and humans makes the completion of a wide variety of complex tasks more efficient and consistent (Walliser et al., 2019). HAT research draws on a variety of domains, ranging from industrial and organizational psychology (e.g., Burke et al., 2006) to team dynamics (e.g., Butler, 1999), to intelligent team tutoring systems (e.g., Gilbert et al., 2017) to human factors (e.g., Parasuraman et al., 2000). One of the key challenges in team research is assessing the team: How did Team A perform on the task? How does Team A compare with Team B? How does Team A’s performance on one task predict its performance on a different task? These questions are complex because the answers stem from a combination of individual team member’s task-related skills, their teamwork skills, and the team’s experience working together (Salas et al., 2007). This challenge becomes further complicated with HATs, due to the fact that a person with good teamwork skills may not have an appropriate mental model of an intelligent agent or know how to read the agent’s cues about its current goals and reasoning process (Chen & Barnes, 2014).

This research explores the process of creating useful team metrics for assessment for a particular team testbed called MazeWorld. Some team-related testbeds have been developed, including BlocksWorld (Johnson et al., 2009), Minecraft for urban search and rescue (Bartlett & Cooke, 2015), military surveillance (Ostrander, Bonner, et al., 2019), and the Team Multiple Errands Task (Walton et al., 2015). However, studying HAT has been a challenge because of the difficulty in developing standardized metrics to capture teaming skills from the kind of observable data that is measured (Ostrander, Gilbert, et al., 2019).

To compare the suitability of these previous testbeds for future research, the authors created the requirements criteria listed in Table 1. The requirements were developed by first identifying four key elements in the design of a testbed: system capabilities, agent properties, task properties, and measures. In each of these categories, specific requirements touched on the level of scalability, flexibility, and feasibility of design elements within these four aspects. Due to the previous testbeds not supporting these as completely as desired, particularly in the areas of scalability, repeatability, and ability to capture desired metrics, the authors designed their own testbed, MazeWorld. This testbed includes generalizable metrics that allow for the systematic exploration of different teaming dimensions that are comparable across studies.

Inspired by efforts to map measurable behavioral markers to constructs that are desired, e.g., team communication (Sottileare et al., 2018), the authors used the Framework for Automated Team Evaluation (FATE) (Figure 1) (Ostrander, Gilbert, et al., 2019) to create team and individual metrics for MazeWorld. The authors suggest that the example process documented by this paper can offer guidance to the more general task of creating team metrics.

Table 1: Testbed Requirements

<table>
<thead>
<tr>
<th>System Requirements</th>
<th>Agent Requirements</th>
<th>Task Requirement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Presence of Tutorial Mode</td>
<td>Agent Can Take Action</td>
<td>Scalable Level of Task Difficulty</td>
</tr>
<tr>
<td>Automated Logging of Measures</td>
<td>Variable Level of Automation</td>
<td>Role Allocation Among Team Members and Agent</td>
</tr>
<tr>
<td>Automated Behavior Pattern Logging (clicks, timing, etc.)</td>
<td>Scalable Agent Avatar Qualities</td>
<td>Simultaneous Parallel Actions by Different Team Members</td>
</tr>
<tr>
<td>Able to be modified</td>
<td></td>
<td>Multiple Correct Ways to Complete Task</td>
</tr>
<tr>
<td>Tools to Process raw data into testable data</td>
<td>Team Member Interdependence</td>
<td>Team is Repeatable (isomorphic)</td>
</tr>
</tbody>
</table>
Rules of the task are easy to learn

<table>
<thead>
<tr>
<th>Measures</th>
</tr>
</thead>
<tbody>
<tr>
<td>Measures Individual Performance Metrics</td>
</tr>
<tr>
<td>Measures Team Performance Metrics</td>
</tr>
<tr>
<td>Measures Team Skills</td>
</tr>
</tbody>
</table>

Figure 1: The Framework for Automated Team Evaluation (FATE), adapted from (Ostrander et al., 2019).

The next section describes the design of the testbed. Then two notional MazeWorld team experiences demonstrate the ability of the measures to capture differences in team behavior. Finally, discussion of the future development of the system is presented.

METHODS

Task Design

A coordinated and search task in a maze was chosen as baseline because it could be scaled and manipulated to meet the requirements in Table 1. Automated logging of keystrokes and placement on the map allows for the identification of trends in efficiency and movement. The maze is easily scalable in size and presents the opportunity of being learnable (the same maze every time), or not learnable (unique every time). Any role may be played by an agent or a human. The task, or game mechanic, was designed to allow parallel actions and subgoal assignments. The task was designed such that each role was dependent on another, increasing the amount of teamwork necessary. Rules of the task were designed to be straightforward and easy to learn. Lastly, due to the nature of the task and subgoals, many metrics for evaluating individual and team skills were facilitated by being able to analyze the communication via text chat.

The MazeWorld Task

MazeWorld is a team coin search within a 3D maze environment. Some participants have a first-person experience in the maze, while others have a bird’s-eye view. The team goal is to collect an allotted number of coins as quickly as possible. This is a role-based task for 3-5 players who take on one of three roles: one Tactical player (bird’s-eye view), one Explorer player (in the maze), and one-three Collectors (in the maze). In the case of multiple collectors, each collector and their coins have a designated color. The task is complete when each collector has brought a specific number of coins to the central home station.

The task takes place in a 3D maze (Figure 2) split into 4 quadrants, which can be scaled to meet the desired level of difficulty of the task. Players communicate via a chat window. Each player’s location is automatically updated in the text chat window when the player changes quadrants. The role of the player determines the visibility of the maze in a top down minimap. The Collectors and Explorers can only see where they have been, and the Tactical can only view one quadrant at a time; the rest of the map is blacked out.

Figure 2: An aerial view of an example of MazeWorld

All players and coins are randomly located in the maze, except for the Explorer who starts in the center of the maze, and the Tactical who has no embodiment and remains at birdseye view throughout to guide others. The Tactical must also mark the coins on the map, as they are gray to the Tactical to begin with, as the task progresses to represent the color of the Collector that is responsible for that coin. In addition, the Tactile must communicate with both the Explorer and Collectors to try and be as efficient as possible to minimize the time the task takes. The purpose of the Explorer role is to unlock coins throughout the map for the Collector. At the beginning of the task, 50% of the coins are visible to the Collector. The Explorer must search and find the remaining 50% of coins and activate them for the Collectors. This design was created to make the Collector dependent on information provided by the Explorer. The purpose of the Collector role is to collect coins corresponding to their color and return them to home. For a Collector’s sub-task to be completed, they must deliver a predetermined number of coins to home. A Collector may only carry one coin at a time. An agent in this human-agent team can fulfill any role on the team. This would allow for the testing of various different team behaviors and strategies. For instance, a Collector could be programmed to be selfish, do what is best for the team, or strategize about
specific situation, depending on the research goals of the study.

**Communication Log**

The text log used for the testbed was formatted so that the movement of any player between Quadrants is automatically recorded. In addition, all movement errors and communication clarification and messages over 125 characters are recorded (and highlighted) for metrics at the end of the task. An example log of 4 players can be seen in Figure 3.

**MazeWorld 4 Player Chat log**

The timer begins, players are dropped into the map.
T: C1 has dropped in Quadrant 1.
*T moves to Quadrant 2.
T: There are 2 shapes in Quadrant 2.
*T moves to Quadrant 3.
T: C2 has dropped in Quadrant 3.
*C moves to Quadrant 4.
T: There are 2 shapes in Quadrant 4.
E: coin identified for Collector 2 found in Quadrant 4.
C2: Acknowledged.
C2: I have collected a coin in Quadrant 2. I am returning the coin to home.
*T move to Quadrant 4.
Movement error recorded from E in Quadrant 4.
T: E do you need any guidance?
E: No thank you.
C1: I’m lost can someone help?
Request for clarification recorded from C1 in Home
T: C1 there is a coin nearby in Quadrant 2 if you go out of home and go up in Quadrant 2 it is close to the outer rim of the map as well as next to Quadrant 1.
Communication from Tactical exceeded 125 characters
C1: Ok.
E: coin identified for Collector 1 found in Quadrant 4.
C1: Acknowledged
C1: I have collected a coin in Quadrant 2. I am returning the coin to home.
C2: I have delivered my first coin home. I am going to Quadrant 4.
C1: I have delivered my first coin home. I am going to Quadrant 4.
C1: C2 left Quadrant 4.
Negative communication detected from C1.
C2: I have collected a coin in Quadrant 4. I am returning the coin to home.
C2: I have delivered my second coin home. I am going to explore the maze.
C1: I have collected a coin in Quadrant 4. I am returning the coin to home.
C1: I have delivered my second coin home.
T: The mission as a whole has been completed. Well done team!

**Figure 3: 4-player example of a communications transcript**

**Metrics**

Metrics for MazeWorld include task performance and team performance. Task performance is broken down into performance of the individual roles, and performance of the team (Table 2). High scoring metrics indicate a positive impact. Whether a metric increases or decreases the overall score is displayed in the score column by an up or down arrow, respectively. Team metrics are broken down into communication, trust, workload, and situational awareness (Table 3).

**Table 2: Task performance metrics**

<table>
<thead>
<tr>
<th>Role</th>
<th>Dependent Variable</th>
<th>Definition of Metric</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tactical</td>
<td>Coins Identified</td>
<td>Amount of coins identified/ total coins in maze</td>
<td>↑</td>
</tr>
<tr>
<td>Explorer</td>
<td>Coins identified</td>
<td>Amount of coins identified/ 1/2 the total coins within maze</td>
<td>↓</td>
</tr>
<tr>
<td>Collector</td>
<td>Number of movement errors</td>
<td>When left or right turn keys exceed 4 consecutive clicks (ie 360°) Keystrokes</td>
<td>↓</td>
</tr>
<tr>
<td>Team</td>
<td>Time</td>
<td>Time to complete task in seconds</td>
<td>↓</td>
</tr>
<tr>
<td>Explorer</td>
<td>Maze explored</td>
<td>Amount of map explored as a percent</td>
<td>↑</td>
</tr>
</tbody>
</table>

**Table 3: Team metrics designed for task**

<table>
<thead>
<tr>
<th>Role</th>
<th>Dependent Variable</th>
<th>Definition of Metric</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Team Communication</td>
<td>Requests for Clarification</td>
<td>Number of requests for clarification - total movement errors recorded</td>
<td>↓</td>
</tr>
<tr>
<td>Team Trust</td>
<td>Negative Communication</td>
<td>Occurrences of team members using negative language from set array list</td>
<td>↓</td>
</tr>
<tr>
<td>Explorer Workload</td>
<td>Maze explored</td>
<td>Amount of map explored/ total map units</td>
<td>↑</td>
</tr>
<tr>
<td>Collector Workload</td>
<td>Maze explored</td>
<td>Amount of map explored/ total map units</td>
<td>↑</td>
</tr>
<tr>
<td>Tactical Workload</td>
<td>Markers</td>
<td>Number of Locations Markers accurately placed</td>
<td>↑</td>
</tr>
<tr>
<td>Team Situational Awareness</td>
<td>Awareness</td>
<td>Accuracy in recalling number of coins collected midway through the game, (correct number identified - total amount recorded)</td>
<td>↑</td>
</tr>
</tbody>
</table>
RESULTS

A key goal of the MazeWorld testbed development was whether the metrics successfully enabled the measurement of the task skills and team skills. Two examples are provided to demonstrate the ability to capture both good and poor performance and teaming. Multiple pilot trials of MazeWorld were run with the metrics manually calculated to evaluate them; Tables 4 and 5 show example metrics from the first trial with team A and Team B, respectively. For the purpose of this test the teams had no limit on time and therefore were both able to complete the task.

Team A was faster at completing the task and when looking at the workload (coverage of the maze) each player contributed roughly equally. While the communication is a -1, this is due to one communication with a length of over 125 characters. No negative language was used as seen by the zero in the trust column and all participants were actively engaged with the other players according to the situational awareness score. This was measured by pausing the task periodically and asking each player how many coins have been collected (a situation awareness probe). For each number given that is incorrect, they lose that number in the score. This shows that during the task they consciously kept track of everyone’s progress within the task. Overall Team A was the faster team and covered half of the map (the task ends as soon as the last coin is brought home, even if the map is not fully explored).

Team B was slower but not significantly, therefore we cannot determine a better team by time alone. However, workload was varied more within Team B than within Team A, which could be due to lack of communication not accounted for in the communication column. The trust is a -2 based on two inappropriate words used in the communication log. This shows frustration within the task between team members. Lastly, the situational awareness was a -3. This means that the players were not actively engaged with their teammates and did not track the progress of the team as a whole and were more concerned with the individual score. This score was based on a mid-game probe in which team members did not know how many coins other members had found. Overall Team A is a better team, with their situational awareness, communication, and balanced workload. Utilizing these metrics researchers can predict that Team A would perform better on tasks.

Table 4: Team A Metrics

<table>
<thead>
<tr>
<th>Performance Team A</th>
<th>Coins Found</th>
<th>Mistakes</th>
<th>Workload</th>
</tr>
</thead>
<tbody>
<tr>
<td>Individual</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tactical</td>
<td>4/4</td>
<td>N/A</td>
<td>4/4</td>
</tr>
<tr>
<td>Explorer</td>
<td>2/2</td>
<td>1</td>
<td>28/163</td>
</tr>
<tr>
<td>Collector 1</td>
<td>2/2</td>
<td>0</td>
<td>33/163</td>
</tr>
<tr>
<td>Collector 2</td>
<td>2/2</td>
<td>0</td>
<td>39/163</td>
</tr>
</tbody>
</table>

Table 5: Team B Metrics

<table>
<thead>
<tr>
<th>Performance Team B</th>
<th>Coins Found</th>
<th>Mistakes</th>
<th>Workload</th>
</tr>
</thead>
<tbody>
<tr>
<td>Individual</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tactical</td>
<td>4/4</td>
<td>N/A</td>
<td>4/4</td>
</tr>
<tr>
<td>Collector 1</td>
<td>2/2</td>
<td>1</td>
<td>20/163</td>
</tr>
<tr>
<td>Collector 2</td>
<td>2/2</td>
<td>2</td>
<td>45/163</td>
</tr>
</tbody>
</table>

DISCUSSION

The metrics designed for this task we created to give an accurate picture of both individual and team performances. It was difficult when creating the metrics to make sure each one was measuring something different. After applying the metrics, the researchers found that while each piece of the metrics were useful, there are still holes remaining. For example, communication was key in understanding the value of the teamwork, but clarification requests are not equal to movement errors and therefore should not be included in the same metric. In a future version, the communication metric would benefit by having a set of restricted trackable phrases that could be mapped to the value added by each communication. It is also important to note that better team behavior metrics, when accompanied by appropriate feedback, could create better team performances.

CONCLUSION

The need for human-agent teams will grow exponentially in the next few decades. Agents are already living in our homes, linking us together, and interacting with us. It is important to understand what factors influence this relationship and the success and effectiveness of HAT. Our goal was to research the effectiveness of HAT through the use of a flexible, scalable, and generalizable testbed. The researchers analyzed and weighed qualities that an ideal testbed would include. A navigation and search task was developed to fulfill the testbed requirements. With the task, the researchers needed variables that could provide a translation from the task to real characteristics of HAT. This testbed will be matured in order
to enable evaluations of relationships between individuals in a team, communication, cohesion, and the HAT. Future work includes development to automate logging metrics and enrich the raw event information to higher level team metric constructs. The ultimate goal of the testbed is to provide a forum for researchers to systematically test human-agent teaming algorithms, constructs, and human factors issues.

REFERENCES


