



Adam Fouse
Aptima Inc.
Washington DC, USA
stefanola@ru.is



Figure 1: Human AI Interaction via Minecraft

ABSTRACT

In this extended abstract we present the design, development, and evaluation of a Minecraft-based simulated task environment to conduct human and AI teaming research. With the deluge of AI-driven applications and their infiltration into many activities of daily living, it is becoming necessary to look at ways that humans and AI can work together. There is a tremendous research burden associated with accurately evaluating the best practices and trade-offs when humans and AI have to collaborate together in completing critical tasks. Minecraft offers a low-cost alternative as an early investigating tool for researchers to build answers to emerging research questions before significantly investing in human-AI teaming activities in the real world. We demonstrate

<https://doi.org/10.1145/3570945.3607305>

successfully via a simple rule-based AI, insights that could highly influence human-AI teaming activities can be derived to improve practical and viable development of protocols and procedures. Our findings indicate that simulated task environments play a critical role in furthering human AI teaming activities.

CCS CONCEPTS

- **Human-centered computing** → *Human computer interaction (HCI)* → **Collaborative interaction**

KEYWORDS

Human-AI Teaming, Simulated Task Environments, Software Testbed

ACM Reference Format:

Ashish Amresh, Nancy Cooke and Adam Fouse, 2023. A Minecraft Based Simulated Task Environment for Human AI Teaming. In *ACM International Conference on Intelligent Virtual Agents (IVA '23)*, September 19-22, 2023, Wurzburg, Germany. ACM, New York, NY, USA, 3 pages. <https://doi.org/10.1145/3570945.3607305>.

1 INTRODUCTION

Urban search and rescue (USR) refers to operations conducted in collapsed man-made structures. It has been recognized as a useful domain for studying human-AI interaction [2]. Human-AI teaming in the domain of USR is a widely researched area [1], due in part to the complications that arise out of introducing AI into an unpredictable environment such as a collapsed building. A simulated task environment (STE) is one where real-world tasks and cognitive decision-making can be performed and evaluated with a certain degree of precision and the results gathered have a strong correlation with operations in the real world [3]. STEs are ideally suited for USR tasks due to the difficulty and high cost of developing hands-on training exercises. There is potential to further improve the performance of STEs by including AI agents that can assist the human performers within the task. In this study, we present the design of an AI agent that aims to assist three-person human teams and helps improve their task performance during USR missions. Specifically, we are investigating the use of different communication styles for the AI agent and its effect on human teams. The project was divided into two phases, a Wizard of Oz (WoZ) design phase where researchers playing the role of an AI advisor, gave guidance to a team of participants during the experiment. Data collected from this phase was used to develop a rule-based AI agent and further data was collected on how varying communication styles performed with human participants under two distinct conditions: 1) push (where the AI agent provided distinct guidance to each team member) and 2) pull (where the AI agent provided guidance when requested by the team member). Figure 1 shows the major components of the STE, to the right we have the Minecraft game environment, to the top-left we have the top-down view of the map of the collapsed building, and to the bottom-left we have the communication interface for the AI.

While the objective of the experiment for the participants was to save as many trapped victims as possible in the 15-minute missions, this end goal while important is not the focus of the study. Instead, we will evaluate teams by how often they adhered to the advisor's guidance, disregarded the guidance, and asked for further information from the advisor. In a recent study [4], USR missions were shown to be a useful experiment for looking at exploration (the guidance cannot be explicitly stated) vs. exploitation (where the guidance is explicit) behaviors in teams because both are required to achieve success as a team, although the correct balance between the behaviors can be hard to find or negotiate between teammates. As a result, the guidance provided by the AI tries to keep a balance between exploration and exploitation tasks necessary to complete the missions.

2 METHOD AND FINDINGS

The study required three-person teams to enter a Minecraft simulation of a USR scenario with the goal of saving as many victims as possible in 15 minutes. Teams played two missions on the same map, designed to simulate a collapsed office building. Experienced video game players were chosen as participants to ensure they would be able to understand and perform the task.

Participants were able to select between three roles, each having one unique tool. The three roles were 1.) medical specialist with the ability to save victims, 2.) hammer specialist, a slow player with the ability to break through rubble-blocking areas of the map, and 3.) search specialist, a speedy player with the ability to pick up and place victims. Participants could select any combination of these roles, and during the mission, could also switch roles by returning to a starting area on the map. The unique abilities possessed by the roles allowed for teams to respond to the changing environment of the map by changing the composition of roles on their team. Prior to each mission, participants engaged in a 3-minute preplanning session in which they would be shown a top-down view of the map in Zoom and would be given the opportunity to discuss their strategy for the upcoming mission. Participants were prompted to discuss what roles they would begin with, where they would go on the map, and how they would work together as a team.

2.1 WoZ Guidance

Data was gathered from 54 participants (48 Men, 8 Women) separated into 18 teams. 85% of the participants were regular video game players, and 71% of participants were Minecraft players specifically. Prior to participating in the experiment, all participants were briefed on the objectives and rules of the experiment with a video presentation. Following this, all participants engaged in an in-game tutorial allowing them to become familiar with the tools used in the experiment. Zoom was used to conduct the experiment remotely and to allow the participants to interact with each other. Later, the Zoom recording of the Minecraft environment and audio transcripts were used to determine adherence to the advisor's guidance. Data on participant adherence to advice was gathered through Zoom transcripts and video recordings of the Minecraft experiment. Any time a participant tried to follow the advisor's guidance; it was counted as adherence. Due to the changing requirements of this mission, there were several instances in which a participant intended to follow the advisor's guidance initially but had to change their plans en route. Instances in which participants disregard or fail to adhere to the guidance were noted as "failed to adhere", and instances in which the guidance led to a discussion amongst two or more teammates were noted as "prompted discussion." Information-shaping guidance appears to be associated with the highest rates of compliance in participants. This relationship calls for further research, considering there are a few reasons that participants could have preferred information shaping over explicit guidance. It is possible that participants didn't have enough trust in the advisor to follow commands from them, but they may have been more likely to accept that the advisor had information about the state of the map. There is also a difference between being told what to do and being told why you might want to do something. It is worth mentioning that many of the participants in this usability study were experienced Minecraft players. While they weren't familiar with the specific experiment they were participating in, it is possible that they preferred to rely on their own expertise as opposed to following explicit instructions.

2.2 Rule Based Guidance

In the next step, we designed an experiment to identify the effects of how guidance from an AI agent is communicated; impacts compliance and team performance with within-subjects (repeated measures) design; with communication style (push or pull) as the independent variable and compliance; performance (number of victims saved and time); map exploration as a dependent variable. We hypothesize that there would be a difference in the compliance of the teams between AI communication methods. The communication methods were defined as “push” or “pull”. The push condition had the agent send out guidance to a team member when that team member and the environment satisfied certain conditions. The pull condition had the agent only send guidance to a team member when guidance was asked for. There were no modifiers for asked for guidance in that the team member would simply indicate that they would like any guidance available from the agent. Four teams of three players participated in two sessions per week of three games over 4 weeks for 24 game plays. The order of manipulation of the role of the AI agent was counterbalanced within and between groups. The manipulation mode switched halfway through the experiment at the end of week 2. Initially, participants completed a survey for demographic information, plus additional questions about how often and what kinds of video games they played as well as their general feeling towards AI agents. After each session, participants completed an additional survey to gauge their feelings about their team’s performance, their individual performance, and how the AI agent impacted their team.

We examined the differences between the push and pull conditions. Outcomes examined across these differences included the number of victims saved in each trial, the number of guidance messages given by the agent, and the proportion of guidance followed. As expected, we found significantly more guidance messages given in the push condition per trial ($M = 26.43$, $SD = 5.00$) than in the pull condition ($M = 6.84$, $SD = 4.03$). However, there was no difference in the proportion of guidance followed, even though as expected more guidance was followed when more guidance was given.

ACKNOWLEDGMENTS

Funding for the project was provided by a DARPA grant.

REFERENCES

- [1] Cade E Bartlett and Nancy J Cooke. 2015. Human-robot teaming in urban search and rescue. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 59. SAGE Publications Sage CA: Los Angeles, CA, 250–254.
- [2] Jennifer L Burke, Robin R Murphy, Erika Rogers, Vladimir J Lumelsky, and Jean Scholtz. 2004. Final report for the DARPA/NSF interdisciplinary study on human robot interaction. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 34, 2 (2004), 103–112.
- [3] Nancy J Cooke and William F Lawless. 2021. Effective human–artificial intelligence teaming. *Systems Engineering and Artificial Intelligence* (2021), 61–75.
- [4] Dylan A Orth, Michael Buchanan, Ashish Amresh, Cassady Smith, Glenn Lematta, Nancy Cooke, Adam Fouse, and Samantha Dubrow. 2021. Designing for exploration and exploitation in experimental search and rescue scenarios. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 65. SAGE Publications Sage CA: Los Angeles, CA, 720–725.