 doi:10.20944/preprints202011.0352.v1

Article

Novel Human-in-the-Loop (HIL) Simulation Method to Study Synthetic Agents and Standardize Human-Machine Teams (HMT)

Praveen Damacharla [1,]*, Parashar Dhakal [2,], Jyothi Priyanka Bandreddi [2,], Ahmad Y. Javaid [2,]*, Jennie J. Gallimore [3], Colin Elkin [4,], and Vijay K. Devabhaktuni [4,]

1 KineticAI Inc., Crown Point, IN 46307, USA; Praveen@KineticAI.com (P.D.)
2 Electrical Engineering and Computer Science Department, the University of Toledo, Toledo, OH, 43606, USA; Parashar.Dhakal@rockets.utoledo.edu, jyothipriyanka.bandreddi@rockets.utoledo.edu (J.B.), Ahmad.Javaid@Utoledo.edu (A.J.)
3 The College of Technology, Architecture and Applied Engineering, Bowling Green State University, Bowling Green, OH, 43403, USA; jgallim@bgsu.edu (J.G)
4 ECE Department, Purdue University Northwest, Hammond, IN, 46323, USA; cpe@pnw.edu (C.E.), Vjdev@pnw.edu (V.D.)
* Correspondence: Praveen@KineticAI.com, Ahmad.Javaid@Utoledo.edu; Tel.:+1-419-450-0357 (P.D.), +1-419-530-8161 (A.J.)

Featured Application: This work can be used to develop reliable Intelligent Systems that have the capabilities to integrate in human-machine teams in dealing with time sensitive situations.

Abstract: This work presents a multi-year study conducted at the University of Toledo, aimed at improving human-machine teaming (HMT) methods and technologies. With the advent of artificial intelligence (AI) into 21st-century machines, collaboration between humans and machines has become highly complicated for real-time applications. The penetration of intelligent and synthetic assistants (IA/SA) in virtually every field has opened up a path to the area of HMT. When it comes to crucial tasks such as patient treatment/care, industrial production, and defense, the use of non-standardized HMT technologies may pose a risk to human lives and billions of taxpayer dollars. A thorough literature survey revealed that there are not many established standards or benchmarks for HMT. In this paper, we propose a method to design an HMT based on a generalized architecture. This design includes the development of an intelligent collaborative system and the human team. Followed by identification of processes and metrics to test and validate the proposed model, we present a novel human-in-the-loop (HIL) simulation method. The effectiveness of this method is demonstrated using two controlled HMT scenarios: emergency care provider (ECP) training, and patient treatment by an experienced medic. Both scenarios include humans processing visual data and performing actions that represent real-world applications while responding to a Voice-Based Synthetic Assistant (VBSA) as a collaborator that keeps track of actions. The impact of various machines, humans, and HMT parameters is presented from the perspective of performance, rules, roles, and operational limitations. The proposed HIL method was found to assist in standardization studies in the pursuit of HMT benchmarking for critical applications. Finally, we present guidelines for designing and benchmarking HMTs based on the case studies' result analysis.

Keywords: artificial agents; human factors; human-machine teaming; metrics; synthetic agents

1. Introduction

The usage of intelligent agents (IA) in everyday life has been increasing immensely. This includes the domains of medicine, military, traffic, and industry and being widespread in many other fields. The U.S. military has keenly started to show its interest in artificial intelligence (AI) and augmented reality

 doi:10.20944/preprints202011.0352.v1

(AR) to improve their medical modeling and simulation (MMS) programs that train doctors, nurses, and first respondents in handling real-world situations. The MMS technology also helps military medical professionals face numerous medical concerns, whether in a military treatment facility or on the battleground [1]. Within the medical field includes the adaption of IA in the diagnosis of Cardiac disorders [2]. Similarly, IA has also been used for the assessment of traffic conditions on highways [3]. In addition, [4,5] describe the applications of IA in multiple types of apparel industries.

IA, often referred to as artificial agents, alludes to a self-ruling entity that evaluates a task like an individual, firm, machine, or program. It can also perform assigned activity autonomously over some time using pieces of given relevant information. Though they might range from being simple or highly complex, they help keen operators learn or utilize information while accomplishing their objectives [6]. Some examples of such agents are Amazon Alexa, Apple Siri, Microsoft Cortana, Google Assistant, and robotic agents. Moreover, IA can be classified into four categories based on their capacity and level of accessibility [7].

1. Simple Reflex Agents: These agents act based on the current inputs while ignoring past data. They work based on the condition-action rule, which relates a condition to action and works only when the environment is fully accessible.
2. Model-Based Reflex Agents: These agents keep track of the perception-history-based internal state and work even when the environment is partially accessible. The current state is stored within the agent and maintains an imprecise/predicted model of the environment's inaccessible sections.
3. Goal-Based Reflex Agents: An extension of Model-Based Reflex Agents, these agents make goal-based decisions. Each activity is intended to decrease the difference between the current state and the goal, thereby permitting the agent to pick the path leading to the goal from different potential decision paths/outcomes. The decision-making strategy can be modified, making these agents adaptive.
4. Utility-Based Agents: These agents are capable of recognizing both the goal and system state. They are also capable of characterizing a proportion of how alluring the state is.i.e., the utility of a state. An objective utility-based agent picks the activity that boosts the utility of the activity results, i.e., what the agent hopes to infer, by and large, given the probabilities and utilities of every possible outcome.

Synthetic Assistants (SA) are one such type of IA that are intended to give people physical help, either by participation or by performing tasks. They can be either software, hardware, or a blend of both that ideally supplements people's capacities in complex errands [8]. SA is most commonly used to aid people with physical disabilities at the workplace, such as monitoring those with impaired vision and facilitating physically demanding tasks. They exist within the scope of explicit models, yet they are regularly intended to be stylishly satisfying to make them attractive to users. There are many advancements in SA categories, such as VBSA, that can be applied to different applications in various emerging fields such as powered predictions, email, and navigation. In addition, there are also examples of failures in the real-life implementation of automated machines that potentially contribute to threats to human lives [4]. Thus, a careful collaboration between humans and machines is needed to overcome limitations and make efficient use of technology.

1.1. Human-machine Teaming (HMT)

More people will speak to a voice assistant than to their partners in the next three years; according to a UN report [9], this shows how IAs will be a part of people's everyday lives. 85% of Americans use at least one artificially intelligent product, and the global use of AI is expected to reach 1.8 billion by 2021. Factors such as growth in discrete industries, high adoption of smart manufacturing practices, and increasing demand for advanced software solutions in high-risk environments are propelling the growth of AI and, by extension, IAs and SAs [9]. This situation creates a strong need to understand how SA and their counterparts affect human lives around them and how they affect overall job performance.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 November 2020 doi:10.20944/preprints202011.0352.v1

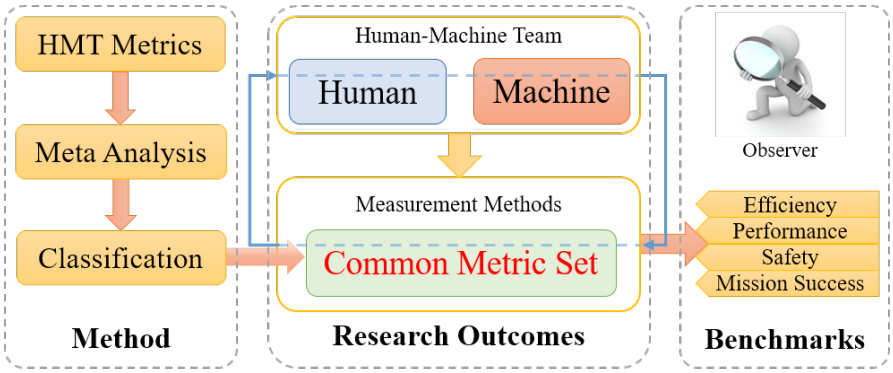

Figure 1. A graphical representation of the proposed novel human-in-loop (HIL) simulation for HMT standardization (reproduced from [5]).

The advancement of AI systems has brought enormous changes to the system workflow and can be evaluated by factors such as system performance, response time, and data rendering. Despite such benefits, there are some challenges, such as consistency and safety issues. One solution to overcome these challenges is human-machine teaming (HMT), which involves strategic cooperation between humans and machines to achieve desired results [10]. HMT can be understood as a combination of cognitive, computer, robotic, and manual resources to maximize team performance [4].

In the HMT field, the composition of strict rules and regulations is critically important when involving humans in tests, as the risk factor is high. HMT could produce incremental safety and efficiency improvements that could be lower than expected unless adequately designed around human factors such as workload and situational awareness. An improperly constructed HMT could create new risks and hazards that threaten a team's safety and efficiency. Therefore, it is essential to measure the impact of HMT on team members and system outcomes. Evaluations of HMT designs necessitate methods and metrics for assessing effectiveness and efficiency, yet such metrics are currently lacking. Thus, HMT's effectiveness must be measured across several dimensions, such as mission effectiveness, machine behavior, human performance, and human-machine collaboration.

1.2. Research Questions and Objectives

Over the past few years, continuous research efforts on human-machine partnering and related methodologies have brought the HMT concept from functional design to reality. The potential benefits of HMT could be limitless and could prompt a new industries in the 21st century. Recently, it has started gaining the recognition it deserves with its multidisciplinary applications. However, HMT is a complex subject and requires rigorous case studies in a controlled environment before its use. Human factor studies are a critical variable of interest in HMT. There is a significant technical advance achieved in autonomous systems (i.e., the machines) through AI, but that is not enough to fully realize an HMT. It requires a different approach to synergize both human and machine performance. In such cases, human performance is a subject of a mental model, task complexity, environment, and interface efficacy, thereby requiring careful assessment. Effective feedback methods are needed that are embedded in every stage of HMT building and operation. This can be achieved by metrics that evaluate an HMT's effectiveness and its agents (e.g., human, machine, and team) on different levels. To this end, we propose several research questions and associated objectives of this study.

Research questions: Although there are ongoing research studies taking place in HMT-based technologies, there is little know-how available in the literature regarding standardization or benchmarking of teaming applications. This is despite the fact that it is critically important to compose rules and roles, validation schema, and approval odds for an HMT application. To achieve this, we pose the following research questions:

1. Could we identify metrics that can be used to measure HMT and its components, and develop a comprehensive measurement method?
2. Could we develop and customize an intelligent agent (IA) that can be used as a machine teammate in HMT?
3. How do we study the developed IA using machine metrics identified in the first research question? Are human-in-the-loop (HIL) simulations (that utilizes the HMT) conducted in a controlled environment would be sufficient?
4. Could performing inferential and exploratory studies on the data collected from HIL simulation studies reveal additional insights that may help standardize or benchmark HMTs for a desired performance level?

Objectives: Testing a hypothesized method and achieving a reliable HMT requires a thorough simulation study on a task-by-task basis. However, there are no common methodologies to study how these systems perform as a team. There are few research publications dedicated to study the performance of a HMT in a single task such as navigation in crowded areas or navigation in stressful situations. Hence, we established the following objectives for this research study to test the above hypothesized research process:

1. Perform a detailed study and meta-analysis of all the available metrics in the literature over human and machine and bring them together under one unique platform. Using the data collected, classify metrics into application and characteristic groups. Using identified groups, establish common sets of metrics that can be used to measure synthetic-assistant-based HMT.
2. Develop a detailed architecture of SA. An integrated SA with an identified interface (e.g. voice) results in a voice-based synthetic assistant (VBSA). In the overall process of HMT architecture development, this serves as a prerequisite to the HMT design.
3. Propose a generalized architecture followed by components and fundamental blocks involved in teaming.
4. Develop a VBSA to be used in HMT that can help ECP in training medical procedures and treating patients in real-time controlled environments. Perform HIL simulation studies that can be used in measuring HMT performance in accomplishing identified tasks.
5. Use identified metrics and established common metrics set in measuring HMT in a Hypertext Markup Language (HTML) simulation, and normalize obtained results that can be used for statistical analysis.
6. Identify the different operational limitations as well as the generalized methodologies to analyze a set of applications in correlation with a specific HMT.
7. Use statistical results and identify limitations as evidence to establish benchmarking for the VBSA-based HMT or to provide guidelines to design future VBSA and HMT, and benchmark them based on lessons learned though this research process.

1.3. Research Paper Outline

This research work has provided insights on benchmarking mechanisms in HMT concepts through simulation studies of VBSA-based human-machine teams. We begin with a literature review and concludes with guidelines to designed VBSA and HMT, and guidelines for performance and benchmarking measuring mechanisms. The following outline has been visually presented in Figure 2. Section 2 begins with background information related to the evolution of artificial intelligent systems and corresponding setbacks. Section 3 elaborates the design and development of VBSA and associated HMT architecture. Consequently, we discuss the proposed HIL HMT simulation implementation using the developed VBSA on two identified test cases. Section 4 presents analytical results obtained through HIL simulation with detailed analysis using HMT metrics. Section 5 provides guidelines on developing new Synthetic Assistant technologies, architecting SA based HMT, and benchmarking VBSA based HMT. This is followed by discussion, conclusions, and future work in Section 6.

 doi:10.20944/preprints202011.0352.v1

1. Introduction	HMT introduction and simulation requirements
2. Background and Literature Review	Understand prior work in designing and benchmarking HMTs
3. Novel HMT Design	Technology/machine teammate in an HMT
4. Results & Statistical Analysis	HMT performance measure, statistical analysis to gain additional insights
5. Standardization Guidelines	Propose HMT design and standardization guidelines
6. Conclusion	Concluding remarks and discussion

Figure 2. Paper outline

2. Background and Literature Review

Even though the term HMT has been coined in recent years, the concept of computers or machines assisting humans can be traced back to World War II. During this period, the French army's armored tanks were much more sophisticated in performance and maneuverability than those of the Germans. However, unlike the Germans, the French tanks had no radios to share information among tanks, which gave a distinct advantage to the Germans over French tanks [11]. This example shows the importance of interaction or communication in the areas of HMT. While no actual teaming was involved between tanks and humans, these applications laid the foundation for the field of HMT. In the assumed symbiotic partnership [12–14], individuals set the goals, establish the conditions, define the criteria, and conduct the evaluations. Simultaneously, computing machines do the pattern work to prepare the way for acumens and conclusions in technical and scientific studies. The age of shared control began in 1963 with Goertz's proposal for handling radioactive material [15]. In shared control, the computer assists in the task by predicting the operator's intent [16]. Later in 2004, [17] proposed methods for bi-directional information flow in the HMI field, officially commenced HMT research.

The invention of the idea of an HMT can be attributed to the game of advanced chess. For instance, in 2005, two amateur chess players, together with three PCs, won a chess tournament against a field of supercomputers and grandmasters [18]. More recently, a current HMT application of notable significance is the F-35 Joint Strike Fighter, which serves as a fighter plane and a mobile sensor kit that collects, correlates, and analyzes an immense amount of data and displays insights on the pilot's helmet [19]. In fact, the military as a whole, has begun to move towards HMT. For instance, the Army plans to fly Apache helicopters and Grey Eagle drones together such that the drones receive commands from the pilot of the aircraft during a mission [20].

Historically, different modalities have also been used by different researchers that play a crucial role in exploring successful HMT implementation. Researchers from [21] used a model-based approach for cognitive assistance to track pilots' changing demands in a dynamic situation. They used flight deck interactions and EEG recordings to find individual pilot's behavior in response to flight deck alerts. Whereas researchers from [22] created a human-robot team where humans had their vehicle to pilot and worked together with several autonomous vehicles to accomplish a common goal in a marine environment. In their work, they planned to use the human operator's physiological measurements to help autonomous teammate vary their communication and levels of control based on the operator's cognitive load. In contrary to others, [23] in their work used image descriptors to create a successful HMT.

In this section state-of-the-art works in HMTs that include currently published architectures, and interfacing/communication methodologies has been reviewed. Furthermore, we also discuss the published HMT simulation techniques and the available benchmarking methods in the HMT area. This literature review will help us to establish clear objectives for this research.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 November 2020 doi:10.20944/preprints202011.0352.v1

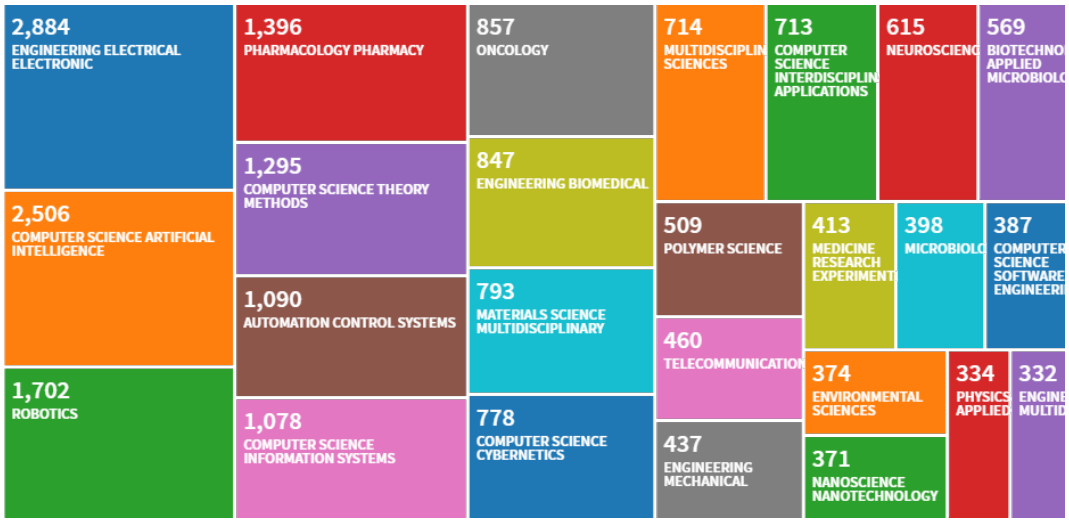

Figure 3. Visualization of research publications in the field of HMT from 2008 to 2020 as per web of sciences ©2020 Clarivate (reproduced from [5]).

2.1. Human Autonomy Teamwork (HAT)

Similar to HMT, Human-autonomy teamwork (HAT) is a process where humans work independently towards a common goal, as does the autonomous agents, but as a team. Here, autonomous agents take a unique role or task and work independently with human team members to achieve a shared objective. During the 90s, HAT was a concept related to autonomous agents playing roles as genuine team players [24]. However, most recently, HAT has been frequently used in association with highly intelligent agents based on AI, machine learning, and cognitive modeling that are as competent as humans. And such work has directed the implementation of human-autonomy collaboration from concept to practice [25–29].

A lot of work during the 90s and recently have been done on the HAT. Researchers from [30] believe that for the HAT to be synchronous, the agent must communicate effectively with human team members. To study the HAT process, they carried out an experiment simulating a synthetic unmanned aerial system environment, where teammates were required to take good photos of critical target waypoint by interacting with each other using a text message. In their work, they defined a team of three. Two were humans acted as navigator and photographer and one was a synthetic teammate who acted as a pilot and the experiment was developed using the ACT-R cognitive modeling architecture. At the beginning of the experiment, human teammates were informed that one of their teammates was synthetic, and to communicate with it, they had to send the message in a very constructive way with no mistake. During the evaluation, they found that the performance score was lower for the synthetic team than that of the other two teams and concluded that the synchrony's quality and effectiveness is more important than the quantity or frequency of synchrony for HAT to be more successful. They also pointed out that effective communication and effective coordination among the team are crucial for effective team synchrony in a HAT.

The researchers from [31] proposed to use design patterns to reduce commonly occurring HAT situations. They investigated the feasibility of reduced crew operations (RCO) for transport category aircraft in their work. The process involved three fundamental human roles: the pilot on board (POB), the dispatcher, and a ground pilot. The POB served as a captain and was responsible for managing automation and ground support. Onboard automation helped POB with flight management system input, assisting with checklists, and validating inputs. While ground-based automation assisted dispatcher with the task such as preflight planning, monitoring aircraft positions, and enroots reroutes. The dispatcher also helped POB's with reducing their workload. Similarly, automation helped the dispatcher create preflight briefings, flight path monitoring, and optimizing reroutes. Automation was

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 November 2020 doi:10.20944/preprints202011.0352.v1

Table 1. Comparison of all six methods in inclusion of required components

Functional Block/ Framework	HMI	Information	System State Control	Mission Planning & Goal recognition	Dynamic Task Allocation
Pilot's Associate	Primitive, no HMT support	Primitive, no HMT support	No	Not available	Primitive
3 Tier Intelligent Framework	Primitive, no HMT support	Real-time monitoring	Yes	Human-based	Machine-based
Crew Assistant	Moderate, no HMT support	Real-time monitoring	No	Human-based	Machine-based
Human-machine Co-op Framework	Theoretically supports HMT	Real-time monitoring	Yes (in theory)	Machine-based	Machine-based
Intelligent Control Framework	Primitive, working model	Real-time monitoring	Yes	Human-based	Machine-based
Coactive Design	Primitive, working model	Real-time monitoring	Yes	Goal-negotiation included	Machine-based

also responsible for monitoring all flights and alerting the dispatcher for any assistance. Besides, the ground pilot also had remote access to fly the aircraft as needed. During this process, it was analyzed that defining design patterns help describe and prescribe human-autonomy relationships.

Another researcher [32] in their study analyzed the role of interaction in Team Situation Awareness (TSA) and team performance to understand HAT. Initially, they analyzed team verbal behaviors such as pushing and pulling information across conditions of human-autonomy teams and human-human teams and further analyzed their relationship with TSA and team performance using Growth Curve Modeling. Good teamwork was supposed to be predicting the needs of teammates and then push information before requested. In their experimental task, participants were instructed to push information to others and master the specific timing of sharing of information to the intended recipient. The study indicated that pushing information was positively associated with TSA and team performance. It was also found that human-autonomy teams had lower levels of both pushing and pulling information than all-human teams. Through the study, the authors concluded that the anticipation of other team member behaviors and information requirements in human-anatomy teams is important for effective TSA and team performance.

2.2. Popular Architecture using HMT Framework

The architecture of the system must be approved at an early stage of the project in order to avoid pricey variations later. Besides, it is important to build the base for a practical HMT architecture by defining the basic essential blocks. To accomplish this, we surveyed 27 frameworks and shortlisted six, based on the presence of essential blocks. We then adapted them to state-of-the-art technology and decided nine essential blocks as a basic HMT framework, namely human-machine interaction (HMI), information, system state control, arbitration, goal recognition and mission planning, dynamic task allocation, rules and roles, validation and verification, and training [13,33]. The different types of architectures that we studied are Pilot's Associate Architecture, 3 Tier Intelligent Framework, Crew Assistant, Human Machine Co-operation framework, Intelligent Control Framework, and Coactive Design Framework. A tabulated analysis of the six major frameworks studied is presented in Table 1, with functional blocks representing how they achieved functions and automation methods.

2.3. Popular HMT Simulation Methods

The teaming concept of human and machine collaboration is an emerging field, and there is a substantial scope and necessity for simulation studies, owing to the intricacies involved in the variables of interest. There is not much information available in the literature that is directly related to simulation studies aimed at evaluation of human and machine team efficacy. Instead, simulation studies can be

 doi:10.20944/preprints202011.0352.v1

found on individual components of HMT in terms of interaction. Following are some of the prominent human-machine systems' software-based simulation methods.

1. Immersive learning simulation: This method is the integration of virtual, augmented, and mixed reality blended with a physical environment aimed at enhanced learning and training. The immersive simulation finds its application in classroom teaching, patient diagnosis, virtual gaming, drone flight control, and computer vision. Immersive learning primarily involves the user's attention and makes the virtual models indistinguishable from the real world. In this process, the user avails head mount displays or stereoscopic projection systems for simulations. With the human user as the subject of simulation, the user's inputs garnered in the process are evaluated as per the task-demands, relayed in conjunction with feedback [34].
2. Modular open robots simulation engine (MORSE): This method is an open-source robotic simulator with a component-based architecture aimed at simulating sensors and actuators of robots [35]. The objective behind modular simulation is to evaluate various levels of the robotic component using the software-in-loop philosophy. The MORSE is capable of multiple simulations of heterogeneous robots (machines) in a fully 3D environment and provides a detailed assessment report.
3. Mixed-initiative system (MIS) plan model to test a team: In this method, the machine generates a plan of action for a specific task, and the user has the authority to intervene and modify the plan if necessary. Generally, learning from demonstration is a technique used in the literature, where a human explicitly teaches a machine a specific skill. However, this does not fulfill a team spirit, which is based on the mutual adaption process for learning coherence in a joint action. In [36], the authors developed a teaming model of human factors engineering with the goal of achieving convergent team behavior during training and task execution. The team model is computationally encoded by extracting the knowledge of the shared mental model of human and machine. In the literature, human-in-the-loop simulations are also known as simulation-based training, and the common areas of application are air traffic control, emergency management, and military operations. For example, consider a UAV team with a single human team member that has the capability of redirection in-flight through transmission of GPS data. The key issue in human supervisory control of a UAV is how a dynamic system that requires constant human re-planning efforts could and should be designed [37].
4. Air traffic control simulation: This method involves the exchange of air traffic management information between stakeholders. The dynamic variation in air traffic management procedures proves the necessity for careful evaluation of traffic control systems through human-in-the-loop simulations. Typically, the implications of potential changes in air traffic management information are initially evaluated through real-time software simulation with the help advanced probabilistic algorithms. This is followed by software-driven results that are used to further assess point-to-point human-in-the-loop connections. Due to the complexity involved in the traffic datasets, the chances of cognitive overload caused by stress factors on a human operator are significantly high, thereby adversely affecting the performance of the air traffic management system. To ease the stress on the human operator, the traditional air traffic control management is replaced by a system-wide information management (SWIM) [38]. SWIM facilitates comfort in human-in-loop simulation by increasing the level of autonomy in air traffic control and introducing immersive techniques. The approach behind human-in-loop studies is to employ an automated process based on machine models and prioritize human intervention towards situations of high uncertainty or high risk. With this in mind, human intervention by decision helps to improve new iterations of the automated models.

2.4. Benchmarking Methods

Benchmarking is imperative for standardizing HMT performance, however, there are very limited relevant information in the literature regarding it. There are several benchmarking methods that have

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 November 2020 doi:10.20944/preprints202011.0352.v1

been developed to assess team efficacy with team members as either machines or humans but not as both together. The most obvious field of benchmarking is seen in human teaming at the industry level. Many companies practice different benchmarking methods to analyze and evaluate operational performance of an organization with a specific mission, the most pronounced of which is known as matrix benchmarking.

Matrix Benchmarking helps in finding lapses in delivery of task execution and indicates areas that require potential improvements. To perform benchmarking of companies, it is important to utilize a generalized categorization approach. The two key variables of a generalized categorization are called cost categories and cost levers [39]. Here, the cost categories are further divided into three parts: payroll, capital and material, whereas cost levers are divided into four parts: production, volume, factor costs, and efficiency. Moreover, these two variables are also utilized to quantify the cost difference between a company's product and its competitors.

The office of personnel management interagency advisory group (IAG) has devised four benchmarking models to evaluate a team performance, as follows [40].

1. Individual team member assessment in a team: In this model, the team member's individual performance is evaluated in a way unrelated to the team objectives. The performance variables of individuals include task complexity, task success, and team member efficacy. The elements of team performance are not considered in this model.
2. Team member contribution to a team setting: Here, the team member's overall performance variables, as well as a team member's contribution to at least one element of the team's mission, are the subject of performance evaluation. In this model, the performance evaluation is conducted from the perspective of individual team member efficacy, while team productivity is not evaluated.
3. Group performance: In this model, the focus is given to performance variables of both the team and individual members. There is at least one element of team group performance related to an individual member in the evaluation process.
4. Group performance (No individual appraisal): The performance is determined at team level, and no individual member appraisal methods are conducted in the due process. This gives insight on overall team productivity when assigned a task.

However, the benchmarking methods at the HMT level require a different set of factors, and to date, no standardization has been proposed. Some of the factors that govern the effectiveness of HMT are discussed in [41]. In fact, there are numerous factors that will determine the extent to which HMT will achieve the desired level of effectiveness when employed. These factors are based on previous research on HMI design and on human-human team performance. Some of the factors that have been discussed in the literature are:

- Automation transparency: The operator must perceive the automation agent's intent and reasoning and be able to elicit additional factors based on the mission objective. In fact, detecting and understanding the status of human team members may be one of the more challenging aspects of autonomous-agent design in HMT.
- Two-way communication: Communications between humans and machines are necessary for establishing and maintaining a common knowledge of task goals, exchanging information, and reporting errors. The advent of natural-language processing incorporated into machines has enhanced communication that matches human capabilities. It is a crucial factor for effective human-autonomy collaborations in pursuit of a common goal.
- Workload: This factor has a direct influence on task execution and team performance. If poorly designed, human interactions with a machine can lead to excessive operator workload due to the difficulty understanding the machine's current intent during a task.
- Situational awareness of human and machines: Humans must have awareness of the tasks, systems, and environments, as well as adequate awareness of the status of the machine and vice versa [42].

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 November 2020 doi:10.20944/preprints202011.0352.v1

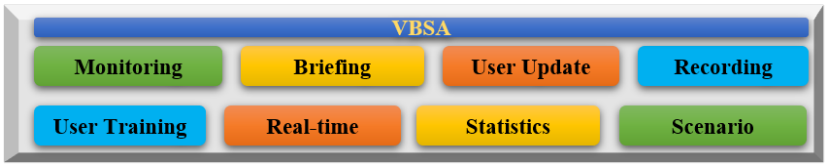

Figure 4. Modes of operations of VBSA (reproduced from [5]).

The desired incremental gains in safety and efficiency produced by HMT could be less than expected unless they are designed properly, taking into account the factors noted above. A poorly designed HMT might even create new risks and hazards that threaten the safety and efficiency of a team. Therefore, it is important to measure the impact of HMT on operator and system outcomes. Evaluations of HMT designs necessitate methods and metrics for assessing effectiveness and efficiency, yet such metrics are currently lacking. The effectiveness of HMT must be measured across several dimensions, such as mission effectiveness, machine behavior, human performance, and human-machine collaboration.

3. Novel Human Machine Team Design

This section introduces VBSA, which is commonly used as a machine in HMT. This section also encompasses motives behind its development as well as its activities, such as statistical monitoring and tracking. This is followed by further details of the proposed HMT architecture, with the major architectural components included. This section also outlines types of metrics used for architectural assessment, such as human, machine, and team metrics.

3.1. VBSA Design

A voice-based SA uses voice interaction as the primary communication mode, and communication with it occurs under the human-based task. The VBSA has been developed to follow major medically approved emergency protocols such as massive hemorrhage, airway blockage, respiration, circulation, hypothermia and eye injury (MARCHE) and airway, breathing, circulation (ABC) [43,44]. This is done to train emergency care providers (ECPs). The developed assistant tracks the training progress of the ECP in real-time and helps the ECP to correct the errors committed, with predefined steps in addition to a timely warning. The SA's secondary role is to provide an ECP with the job details required and to maintain a record of the training results, including the total number of errors, frequent errors and the execution time. After reviewing 29 existing autonomous systems, the architecture of the VBSA has been developed with a combination of a natural language processor (NLP), a speech synthesizer (SS) and primary operation blocks (data pre-processor, scenario calculator, process monitor, result synthesizer and data post-processor). The developed SA technology is integrated into Amazon Web Services (AWS) with respect to language features and micro-services available. [8] states that Alexa Voice Services (AVS) uses both NLP and SS and can be accurately trained for different accents. The database was developed and stored in Amazon DynamoDB, and the primary operation blocks (scenario calculator, process monitor, result synthesizer) were developed in AWS Lambda. Users can access the VBSA using either the Echo smart speaker or the Amazon Alexa mobile application. In our process, the user's performance and the SA were tracked using a dashboard projected on a screen. The errors made during the communication, along with step numbers, were displayed on the dashboard for correction later. A detailed discussion of the tracking process has been discussed in our previous paper [8]. The detailed version of the VBSA architecture and development process has been inspired by our previous research on HMT [8,45–47]. Figure 4 presents modes of operation of the VBSA, while Figure 5 presents the UML diagram of the VBSA architecture depicting the AWS components that were used to develop the cloud-based IA. VBSA developed as cloud based architecture using AWS, according to [48] the cloud architecture are easy to develop, scale, and deploy. However, there will be

 doi:10.20944/preprints202011.0352.v1

a significant delays in the system response based on network and communication protocols which may effects the performance of VBSA.

This VBSA is capable of eight modes of operations, as shown in Figure 4. In the user training phase, the VBSA briefs the trainee regarding the task procedure and proceeds to real-time monitoring of the training session with a recording option if necessary. In conjunction, the VBSA is also capable of providing the user with task updates and runs a scenario upon instruction that includes forecasting statistics. Indeed the VBSA can execute all the modes of operations highlighted in Figure 4 collectively with an option to disable either of them if needed. On the other hand, a observer uses a web-based dashboard, as shown in [8] fig. 4, to monitor the training progress of all the trainees, whereas the administrative staff uses a web-based form as an interface to build new training scenarios. This HTML web based interface plays a key role in collecting interface, error, and performance data from experiments and it acts as repository for all participants data if observers need to go back and see any data from past participants.

Limitations of VBSA rendered by scripted vocabulary had an observable effect on an ECP's interaction with the SA in training that needs further investigation. The possibility of implementing a VBSA in real training is viable only if VBSA is subjected to rapid development for real-time application, verification, and validation of VBSA, and sustainability of deployment.

3.2. HMT Architecture Design

Constructing a generalized system requires a few primary components or building blocks. In our previous work, we have identified six major components of an HMT, with emphasis on architectures, interfaces, and metrics based on various literature [36,49–61]. These are defined in details in our previous work [4].

1. Architectures
2. Interfaces
3. Metrics
4. Roles and rules
5. Team building
6. Verification and Validation (V&V)

These components may be used to construct a generalized HMT architecture. However, the processes of constructing/designing HMT architecture starts with establishing team requirements. The basic requirements of HMT can be defined as follows [5]:

- Machine task. E.g., a VBSA might be required to "assist ECP with error detection and recommend steps to resolve errors."
- Human-related task. E.g., a Human ECP may need to "provide emergency care to the patient and simultaneously report to SA."
- Interface mode. E.g., The communication mode may "prevent any hindrances to ECP's motor skills while treating the patient"

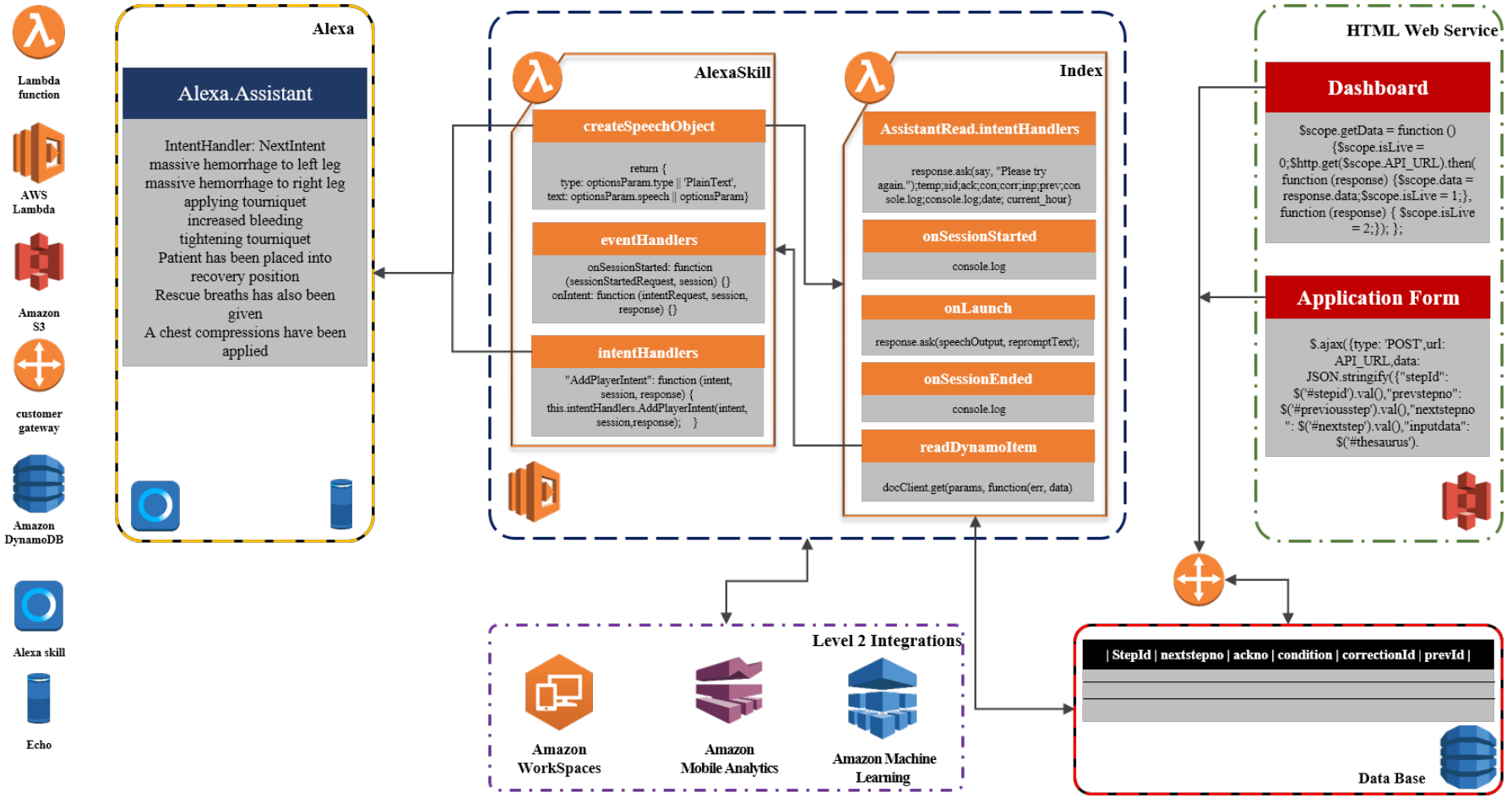

Figure 5. UML diagram depicting the AWS architecture of VBSA (reproduced from [5]).

doi:10.20944/preprints202011.0352.v1

 doi:10.20944/preprints202011.0352.v1

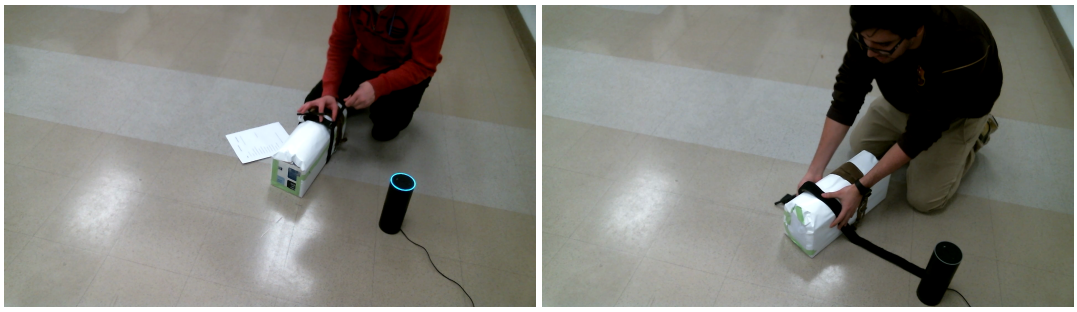

Figure 6. Experimental setup: participants treating model patient leg with massive hemorrhage (reproduced from [8])

Although the foundations of HMT were laid at Defense Advanced Research Projects Agency in 2001 [54], it took five years more for the research community to identify a set of metrics that facilitates a well-organized structure of human robot interactions (HRI). For many metrics, we found close but different descriptions of the same metric, primarily for HMI in human-robot or robot-only swarms [55,62]. Identifying common metrics will allow benchmarking of HMT designs, comparison of findings, and development of evaluation tools. The primary difficulty in defining common metrics is the diverse range of HMT applications. Here, we focus on metrics for all three agents of HMT, i.e. human, machine, and team. The goals of this section are (1) identification and classification of metrics, (2) evaluation of the identified metrics to find common metrics, and (3) proposal of common metrics that can be used in future HMT benchmarking.

Once the HMT is developed, we also need to measure its effectiveness in accomplishing the assigned task. To this end, we compiled a list of metrics through the analysis of 188 works and thorough review of 77 articles among these. We further categorized them into Human, Machine, and Team metrics as part of our survey on common metrics [4].

3.3. Human-in-the-Loop (HIL) Simulation Method

Once the architecture of HMT is finalized and VBSA is set to work in real-time, time-sensitive scenarios, the next step would be the simulation of HMT. To accomplish this, we tested two cases 1) Emergency Care Provider (ECP) training and 2) Medic Treating Patient. Both of these scenarios are time-sensitive as inaction or delay may result in loss of human life. The former case is detailed in [8], and the later case of medic treating patient is detailed in this section. This section also demonstrates steps included in the simulation such as the testing scenario, how the experiment is configured, on what basis the experiment is validated, the categorized list of participants, the procedure of the experiment simulation, and the process in which data is collected and analyzed.

3.3.1. Experiment Case 1: Emergency Care Provider Training

Once the VBSA is developed, it has been tested on a MARCHE protocol scenario, “massive hemorrhage,” in which a patient who is badly wounded with heavy bleeding and needs immediate medical attention by an ECP is treated using a VBSA to stop bleeding. Though VBSA cannot replace a trainer, it enhances the training by providing a sequence of steps to be followed during treatment. A leg manikin model is considered as the patient is treated using an Amazon Echo device that acts as the VBSA interface. Here, the trainer follows the series of actions step-wise, assisted by the Amazon Echo, and helps the patient stop bleeding. This experiment has been conducted using multiple categories of participants, and the data is collected and analyzed based on the survey that is conducted on the participants. [8] provides a detailed explanation of this experiment including data analysis and results.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 November 2020 doi:10.20944/preprints202011.0352.v1

Figure 7. Experimental setup: participants treating manikin patient in Case 2 (reproduced from [5])

3.3.2. Experiment Case 2: Medic Treating Patient

1. Testing Scenario: The testing scenario developed in this case has a major portion in common with the first case study. The scenario explains a situation in which a patient suffers profuse bleeding and a block in the airways and therefore needs immediate medical attention. The task assigned to ECP begins with the cessation of bleeding by proper application of two tourniquets as mentioned in Case Study 1. It also involves assessing a patient's airways, respiration rate, and capillary refill and pulse rate, then concludes with the patient's history collection. The total number of steps composed to accomplish an established scenario is seventeen. Here, the task to assess a patient's airways comprises of seven steps as well as another two steps to diagnose a patient's respiration rate. Meanwhile, the task to assess capillary refill and pulse rate comprises of four steps and another four for collecting the patient's background information. The process protocol is also chosen from military medic first responder training, known as MARCHE.
2. Experimental Setup: A complete human manikin model with a massive hemorrhage on its left leg was considered for the experiment, as shown in Figure 7. In addition, Amazon Echo, based on the VBSA platform, was also utilized in this experiment similarly to Case Study 1 to assist trainees in emergency medical care. The experiment is carried out at the Immersive Simulation Center in a controlled environment at The University of Toledo. As part of the experimental setup briefing, testing and survey rooms were arranged with total isolation. The testing room comprised of a human manikin model, VBSA, ECP, and information dashboard with displays. The purpose of the task execution display is to avoid an additional workload that the participant has to endure by remembering the entire step sequence. The experiment is also video logged with participants' identity discretion, while an independent observer appraises the performance of the trainee and the VBSA. The performance metrics such as task execution, implementation errors, recording errors, and subjective cognitive load of the participant are taken into consideration. Interested readers are encouraged to refer to [5,47] for detailed information on platform selection and software development.
3. Experiment Objectives: In this experiment, the objective of the case study is to evaluate the performance of the medic and the VBSA individually and as a team. As per the established scenario, VBSA assists the ECP in providing medical assistance to the patient. In this process, parameters such as the errors committed by both human and VBSA during task execution, mental workload experienced by an ECP, the VBSA's task monitoring performance, interaction effort between ECP and VBSA, effect of NLP on task success, team cohesion, and time taken to finish the task are put to evaluation. The prime difference between Case Studies 1 and 2 is the participant's medical pre-knowledge in emergency medical care and the absence of a control population set in the experimental process. It is hypothesized that the combination of human and machine team will greatly enhance task execution and capacity handling. In addition, the application of VBSA also reflects on minimization of errors.
4. Participants: As aforementioned in the above section, it was required that the participants have pre-knowledge of the medical process; thus, medical students were chosen for the experiment.

 doi:10.20944/preprints202011.0352.v1

As presented in Table 2, a total of 52 candidates from different branches of medical school at The University of Toledo participated in the experiment. The range of degrees held by the participants varies from high school to Ph.D. Among them, eighteen candidates were male, while the rest were female. Advertisements were posted within the medical college to hire people who volunteered for the experiment. A compensation sum of fifteen dollars, although reasonably not a large amount, was paid to the participants as part of compensation for participation. Due to the candidates' voluntary participation and the absence of an English proficiency test, no screening test was necessary before participation. All the participants were briefed with the help of video aids about testing scenarios prior to the experiments. The briefing included the treatment process and VBSA usage in no more than five minutes to paint an equal mental picture of the situation for all participants.

5. Experimental Procedure: The experiment conducted in this study is similar to that of Case 1 and comprises of two trials. Each trial lasted from 60 to 120 seconds. The scenario comprised of 17 execution steps and began with a hemorrhage treatment performed on a complete manikin followed by diagnosis of three different parameters of a patient's condition, such as breathing airways, respiration rate, and capillary refill, and it concluded with the gathering of patient background information. The experiment was conducted in quiet rooms (e.g. briefing, testing, and survey) under total isolation, and any chance of contact with fellow participants was strictly prohibited in order to avoid human factor bias. The process was as follows:

Table 2. Participants distribution statistics

Category	Count
Total participants	52
Male	18
Female	34
Medicine 2nd year	11
Medicine 3nd year	13
Nursing 2nd year	7
Nursing 3rd year	9
Other medical education background (experts)	12
People with pre knowledge of treatment process	38
People expertise in treatment process	12
People occasional interaction with voice based personal assistant	39
People frequently used voice based personal assistant	13
Midwestern English accents speakers	24
Native English accents speakers	36
Non-native accents speakers	16

   (a) Initially, each and every participant was briefed about the scenario individually with the help of a two minute video. The task entailed hemorrhage treatment and patient diagnosis process-flow with the help of a VBSA. Right after the briefing, the participants were taken to a testing room for task execution.
   (b) During the testing period, the participants were asked to execute total task execution steps on the patient manikin model, which were made available on a dashboard screen.
   (c) Participants began the task with massive hemorrhage treatment similar to that of Case Study 1, then checked a patient's airways and cleared any obstruction by applying chest compression and rescue breath.
   (d) Later, patient respiration diagnosis was performed, and the patient was placed in a tripod position in order to relieve any chronic obstructive pulmonary disease (COPD) detected.
   (e) Adding to that, a capillary refill was made by examining physical appearance (e.g. skin color, nail press). From there, the pulse rate was immediately evaluated.
   (f) Finally, a patient's background information, such as weight, age, gender, and allergies, was gathered. The examiner recorded the testing period as per standard data collection guidelines.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 November 2020 doi:10.20944/preprints202011.0352.v1

(g) After the testing session, the participant proceeded to a survey room for user validation. This consisted of a post-test questionnaire related to user task experience.

The data collected in the due process, illustrated in the section below, was then put to extensive scrutiny to extract results. VBSA Web interface mentioned in section 3.1 plays a key role.

6. Data Collection: The experimental data was collected from time-stamped video recordings logged during the testing scenario and user validation processes conducted after the testing scenario. A real-time video recording was made by an external observer who was responsible for capturing every detail of the task scenario such as VBSA tracking performance, time of execution, errors in executing task sequence steps, and rate of response. The user validation process involved a post-test questionnaire to extract a user's feedback related to task experience. Metrics were collected in this study and were submitted for further subjective analysis. The rate of response, number of interventions, and tracking performance of VBSA were also considered as part of team performance evaluation. The data collected in this due process was evaluated on similar grounds with insights from Section 4.2.9.
7. Data Analysis: As mentioned above, the objective behind this experiment is to assess the performance of an HMT. In this regard, emphasis was placed on performance assessment of the participant and the VBSA, both individually and as a team. This was achieved from the metrics recorded during the testing scenario. The performance score was calculated from metrics using Equation (1), similarly to Section 4.2.8. The formula also incorporated penalty values generated during the experiment.

$$P - score = 10 - (Penalty\ for\ time + \sum_{i=1}^{17}(Penalty\ for\ implementation\ error + Penalty\ for\ recording\ error)) \quad (1)$$

$$Penalty\ for\ time \begin{cases} if\ task\ time\ (t) < 60,\ penalty\ = 0 \\ if\ 60 < t < 120, penalty = (t - 50) * 0.02 \\ if\ t > 120,\ penalty = 10\ (task\ failure) \end{cases}$$

It is also imperative to assess the participant's workload in the process of achieving task goals, as it directly influences the team's task success. The workload endured by the participant while pursuing a team task was divided into six subjective sub-scales: mental demand, physical demand, temporal demand, performance, effort, and frustration. Each of these sub-scales had an exclusive self-rating index ranging from 0 to 100 points at the rate of 5 points per division. The post-testing questionnaire adopted in the user validation process was used to assign self-rating points. In addition, each sub-scale has an individual weight that is multiplied with self-rating points. Finally, the overall workload share was determined using the weighted average of all sub-scales, as shown in Equation (2), which is known as NASA-TLX. In the NASA-TLX formula, MD, PD, TD, performance, effort and frustration are known as sub-scale self-rating indices, whereas W1, W2, W3, W4, W5, and W6 are weights of sub-scales.

In the event that two participant groups were uneven in size to address this challenge for a fair comparison, we used a random selection method to fair comparison groups. In this method, we randomly selected a small sub-group from large group participants. The size of the sub-group was equal to that of the smallest compared group in the statistical analysis. For example, in expert and amateur comparisons, we randomly selected 12 data points from a large pool of a 40-data-point amateur group using the rand function and compared each group to one other.

$$NASA\ TLX\ Score\ (Overall\ workload\ score) = (W1 * MD + W2 * PD + W3 * TD + W4 * Performance + W5 * Effort + W6 * Frustration\ ) / 6 \quad (2)$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 November 2020 doi:10.20944/preprints202011.0352.v1

4. Statistical Results and Analysis

In order to address the fourth research question mentioned in sub-section 1.2, we analyzed data obtained through HIL studies to draw meaningful conclusion and form reasonable HMT guidelines by performing statistical analyses on two sets of data namely performance metrics and survey results in this section. In addition, we also presented results and analysis from the HIL simulations of two cases (i) Paramedic Training of MARCHp Protocol and (ii) Medic treating patient through ABC. First we tried to identify how various groups performed on the HIL experiments and how statistically significant was it. Then we analyzed how parameters such as interface (NLP), machine error, human error affected total outcome/performance of HMT based on these statistical significance. Finally, after analysing the aforementioned parameters we will be proposing few rules to design future VBSA based HMT.

Employing statistical methods such as Welch's t-test, two-tailed Wilcoxon sign ranked test, two-way analysis of variance (ANOVA), and item response theory (IRT) will help understand how different groups were identified in these cases performed compared to overall task success. In addition, we also analyzed parameters, such as team situation awareness, team cognition, human/machine errors, and time of response, that affect team performance and success. Understanding of these factors can help successfully draft guidelines or suggest rule to design a standardized HMT. Moreover, using statistical results obtained in this section will help us formulate and summarize parameters during the standardization of an HMT because those parameters along with appropriate values are necessary to achieve successful HMT formation.

4.1. Case 1 Results and Analysis

Table 3. Two way ANOVA with replication analysis of P-score

Source of Variation	SS	df	MS	F	P-value	F crit
Control	97.4654	5	19.4930	25.7288	1.5130E-19	2.2643
Experiment	2030.9911	1	2030.9911	2680.6986	4.789E-110	3.8936
Control Vs Experiment	5.0054	5	1.0010	1.3213	0.2569	2.2643
	136.3743	180	0.7576			
Total	2269.8362	191				

Table 4. Two way ANOVA with replication analysis of Pscore and execution time

Source of Variation	SS	df	MS	F	P-value	F crit
Control	685.1610	5	137.0322	1.3952	0.0281	2.2643
Experiment	72522.8686	1	72522.86	738.4072	1.3285E-65	3.8936
Control Vs Experiment	1053.4348	5	210.6869	2.1451	0.0421	2.2643
	17678.7493	180	98.2152			
Total	91940.2138	191				

As previously mentioned, Case 1 is discussed extensively in [8], but the analysis of published results regarding Case 1 are limited and preliminary. So, to summarize the findings from [8] analysis, the results from the Welch's t-test and the two-tailed Wilcoxon sign ranked test indicate that the use of a VBSA appears to have no adverse effects on the individual's task execution speed, cognitive load, or stress factor.

However, to further analyze these results we implemented Two way ANOVA method on results obtained through case study 1. The hypothesis is that there is no relation between the experimental

 doi:10.20944/preprints202011.0352.v1

group that uses VBSA in training a ECP, and control group that don't use a VBSA in training a ECP. Tables 3 and 4 are the results of variation between the control group and experimental group with respect to training time and performance. These tests are performed to identify interrelation between the two groups while training progresses. Here, we can observe significant results in Table 4 using p-value.

Based on the results presented above we can see that there is no statistically significance between two groups in their training performance (p-value < 0.005). However, we can see in the columns three controls vs experiment, where we observed statistical significance (p-value > 0.005) between the two groups performance during testing; that means group trained with a VBSA performed significantly better than the group that was not trained with VBSA. Few points to notice from the results of Welch's t-test and the two-tailed Wilcoxon sign ranked test [8], and Two way ANOVA is that the group that was trained with VBSA performed significantly better in their tests and there was no significant effect of performance of ECP student while learning when they were using a VBSA. One conclusion we can draw from case 1 analysis is forming HMT team to train ECP student, or student training in general can improve student overall performance in the field. Second point to observe is even through there is no statistically significant effect on training using VBSA, we can observe some non-native English speaking students may face problem to interact with VBSA, and some student may face problem if have single language structure leading difficulty in using same interaction commands [5,8].

4.2. Case 2 Results and Analysis

Case 2 presents complete results and analysis of the experiment performed. First, we will observe user self reported results, and statistical analysis of these results using the data on NASA-TLX. Second, we are going to analyze statistical results of P-Scores measured in experimentation for every participant, to study situational awareness, individual team member performance, and HMT performance. These analyses will help us validate our research questions 1 to 4 and to draw significant rules to standardize HMTs.

In figure 8 we presented NASA-TLX parameter ranking results for all 52 participants. It can be observed that the participants who used a VBSA as a teammate in treating patients have very low PD, Frustration, and TD and their overall performance also stood high based on weighted participants self score. Figure 9 and table 5 present the data for overall NASA task load score which is weighted and scaled. Based on one one-tailed T-Test on NASA TLX overall performance score for different groups we can observe that there is statistically significant performance difference between native and non-native speakers (p = 0.38 for p-values < 0.05), and for experts and amateurs (p = 0.22 for p < 0.05) who are new to treatment process. This can be interpreted as difficulty with communication with English language and this interpretation of result also validates our case 1 findings. Now when it comes to experts, they already know the process so it takes less time for them to understand VBSA and they are also able to complete experiment task more effectively. This can be concluded as a need for team mates pre-training with VBSA before using it in the field as this pre-training will help form necessary HMTs bonds.

Table 5. One-tailed T-Test analysis on NASA-TLX overall score for different groups

Groups	t	df	p value for p <0.05
Expert and amateur	-0.76772	11	.225409
Native and non-native	-0.29698	15	.384265
Male and female	-3.35829	15	.001073
Medical and nursing	1.29857	10	.101834

If we observe figure 9 for ranked NASA-TLX score for HMT identified groups, it is not possible to draw any significant conclusions from this picture without going to the variance of the data. The data here shows that female participants performance is significantly better compared to male participants as group. The self rated TLX scored are very similar to native and non-native speakers who interacted

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 November 2020 doi:10.20944/preprints202011.0352.v1

with VBSA, but their measured performance showed a significant difference. The conclusions we can draw from these results are that human teammate self analysis of team performance is significantly varying compared to observed measurement.

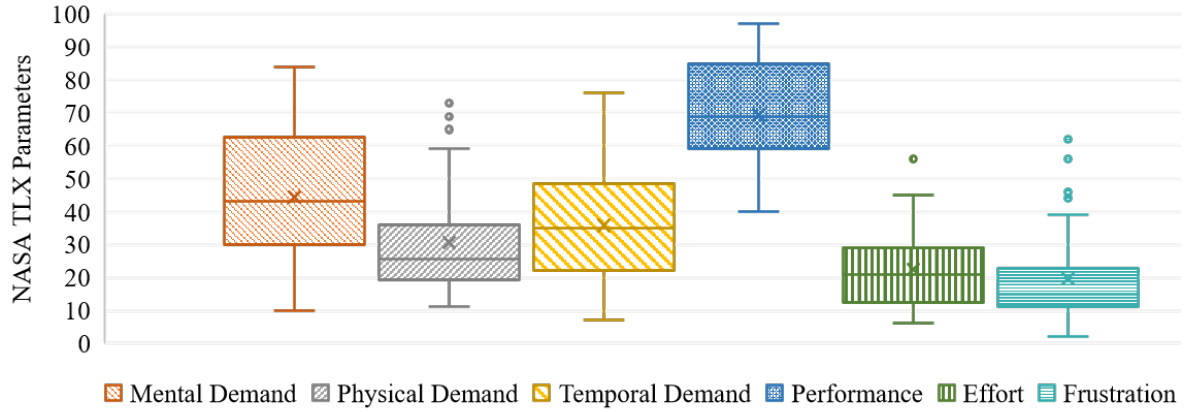

Figure 8. The NASA task load index (NASA-TLX) parameter ranking

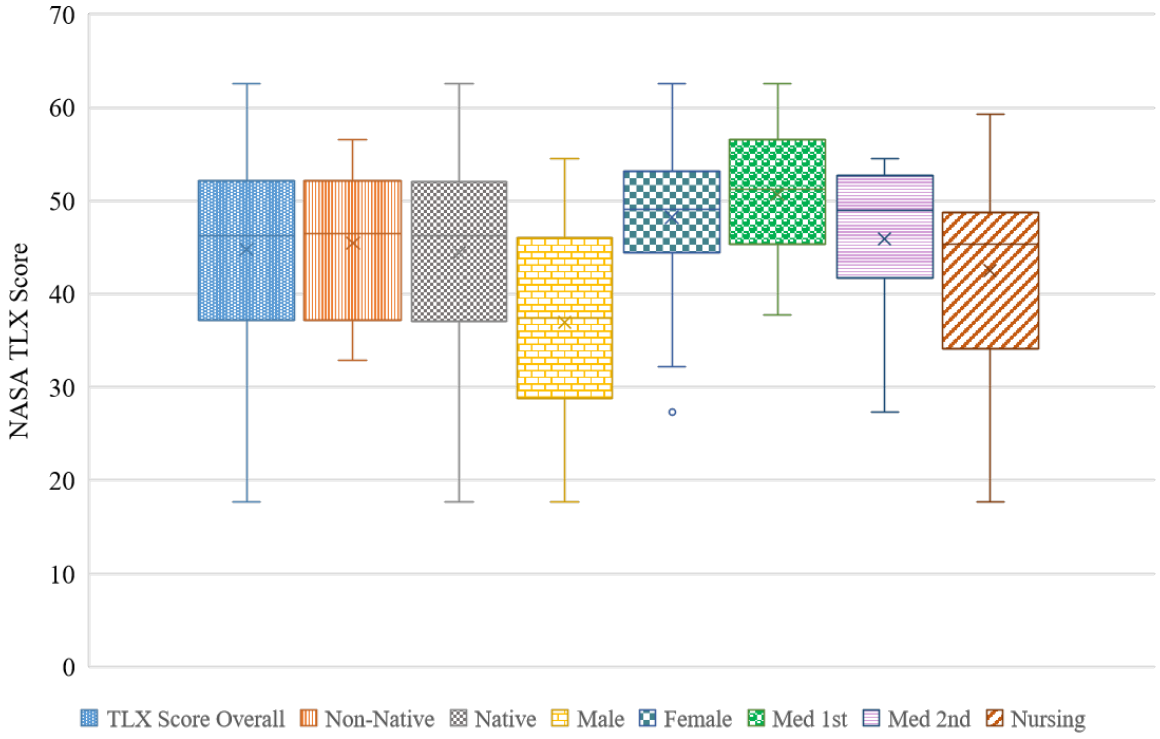

Figure 9. NASA-TLX overall score comparison among HMT identified groups

Performance scores (P-Score) were calculated based on Equations (1) and (2). The P-scores are based on data collected by observers (Task success) and VBSA (Human Error, SA error), which is not self reported. P-score provides on over view of how an HMT performed with respect to task success and how individual team member performed and contributed to that success. P-score can give nuanced information about effect of each parameters of the overall task success. Further analysis for statistical results are presented in Tables 5 through 10. For this analysis, the sample sizes are evened through random selection. In this process, if any group has a large sample size compared to its corresponding group, we used the Excel *rand* function to randomly select data points such that the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 November 2020 doi:10.20944/preprints202011.0352.v1

number of points matches that of the smaller group. Statistical analyses presented in Tables 5 through 9 were tested for p values less than 0.05. If the resultant p value was less than 0.05, then the statistical group performed the test better as an HMT when compared to its counterpart group.

In regards to team performance two separate 8 (conditions) x 2 (iteration) split-plot Analysis of Variance (ANOVA) were conducted to determine whether the team composition (i.e. VBSA - Expert, VBSA-Amateur, VBSA-Native Speaker, VBSA-Non native speaker) differed with respect to their performance and team situation awareness improvement over time (i.e. 2-iteration) refer to table 6, 7, 9, and 10. According to the findings of the first split plot analysis, there was a significant condition main effect [$p<0.05$], while there were no significant interaction effects of condition by iteration [$p = 0.046$ for $p < 0.05$] and no iteration main effect [$p = 0.46$ for $p < 0.05$]. According to the significant condition main effect, the pairwise comparisons indicate that teams VBSA-Expert and VBSA-Native condition performed significantly better than the other three conditions ($p<0.05$). While teams in VBSA-Nursing and VBSA-Medical conditions performed equally ($p = 0.246$), they performed significantly better than the VBSA-Amateur condition ($p<0.0001$; see 7). Overall, these results show that when a trained human that is natural with language teamed up with a VBSA, their overall performance increased. A reason for this outcome can be understood in the context of the major strengths of contemporary NLP models used to develop the VBSA agents. NLP models achieved natural conversational performance across a variety of domains.

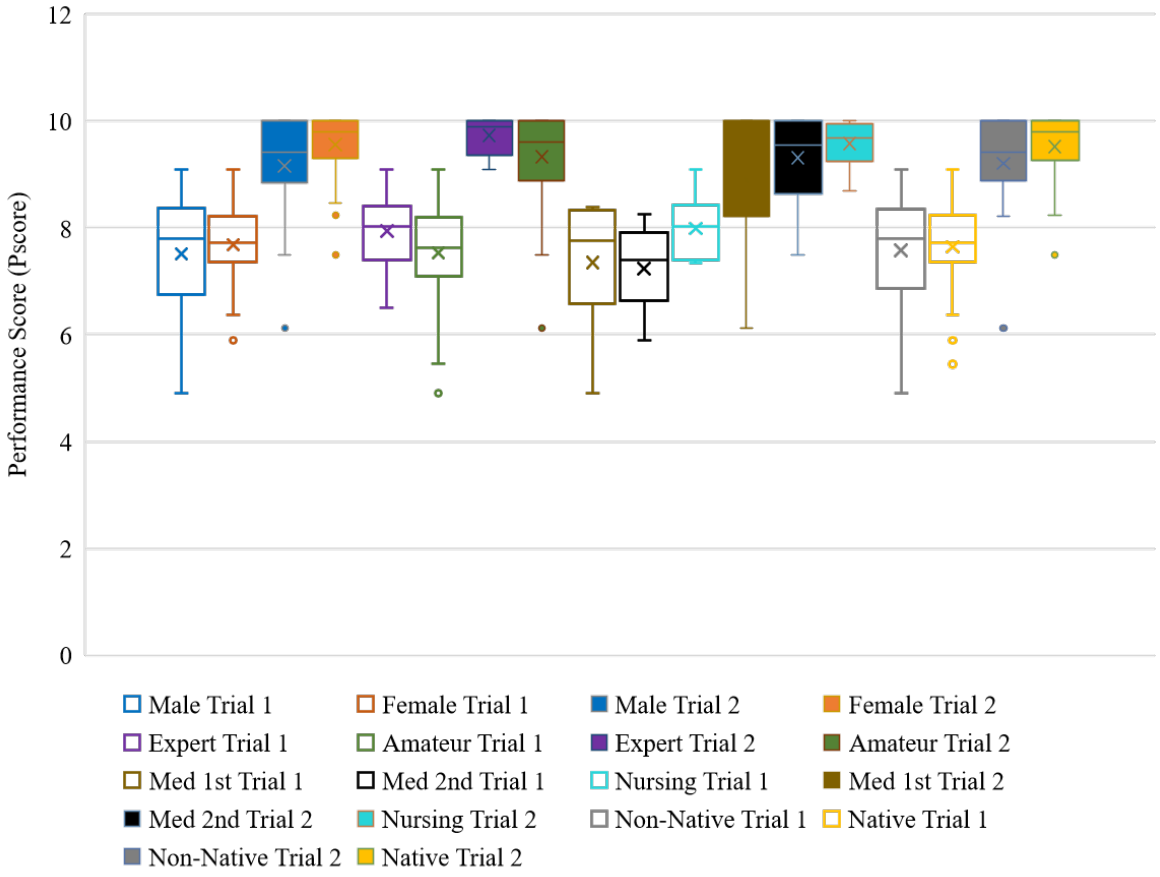

Figure 10. P-score comparison among HMT identified groups

 doi:10.20944/preprints202011.0352.v1

Table 6. Two-tailed T-Test analysis on Trial 2 P-score for different groups

Groups	t	df	p value for p <0.05
Expert and amateur	1.11914	11	.275152
Native and non-native	-2.07168	15	.046989
Male and female	2.05201	15	.047933
Medical and nursing	-1.18088	10	.246633

Table 7. One-Way Repeated ANOVA for P-score between expert and amateur groups

Source	SS	df	MS
Between	45.9933	3	15.3311
Within	38.4264	44	0.8733
Error	28.5213	33	0.8643
F = 17.73856	p <.00001 for p <.05		

Table 8. One-Way Repeated ANOVA for P-score between medical 1st year, 2nd year, and nursing groups

Source	SS	df	MS
Between	2.1113	2	1.0556
Within	24.1031	30	0.8034
Error	10.951	20	0.5475
F = 1.92793	p = 0.171533 for p <0.05		

Table 9. One-way repeated ANOVA for P-score between native and non-native English speakers

Source	SS	df	MS
Between	63.3135	3	21.1045
Within	40.0628	60	0.6677
Error	10.951	20	0.5475
F = 37.96183	p <0.0001 for p <0.05		

Table 10. One-way repeated ANOVA for P-score between male and female groups

Source	SS	df	MS
Between	63.3135	3	21.1045
Within	49.3414	68	0.7256
Error	34.3081	51	0.6727
F = 30.1589	p <0.0001 for p <0.05		

The situational awareness is measured based on two parameters in this test case, those are number of errors a human teammate made in executing a task and depth of involvement from human teammate measure using interface time from figure 11, and 12. A 8x2 split-plot analysis was conducted and tests similar to team performance (score) were ran for team situational awareness. Just as with team score, there was a significant condition main effect [F(2, 30.0159), p<0.0001], while there were no significant interaction effect of condition by iteration [F(12, 159)= 1.49, p = 0.132] and no iteration main effect [F(4, 40), p = 0.043]. According to the significant condition main effect, the VBSA-expert and VBSA-native teams significantly better performance (p<0.0001), which makes sense given how team situational awareness is highly related to team score (higher team situational awareness necessarily means that many resources reached events to complete a given task). Four other conditions did not differ on team

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 November 2020 doi:10.20944/preprints202011.0352.v1

situation awareness: VBSA- non-netive speaker p = 0.37, VBSA-Armature p = 0.69, VBSA-medical p = 0.43; and VBSA-Nursing p=0.54. Overall, these findings make sense given the virtually no gap in team situational awareness between VBSA- (nursing, medical, and amateur) teams.

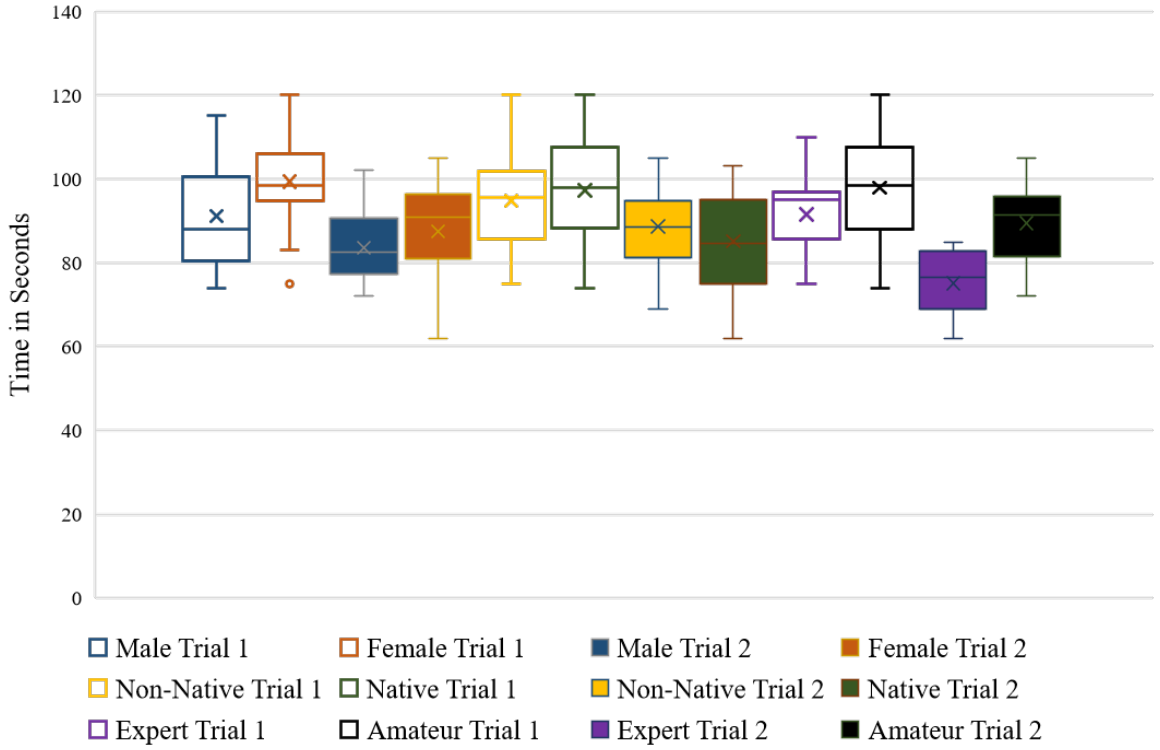

Figure 11. Task completion time for HMT identified groups

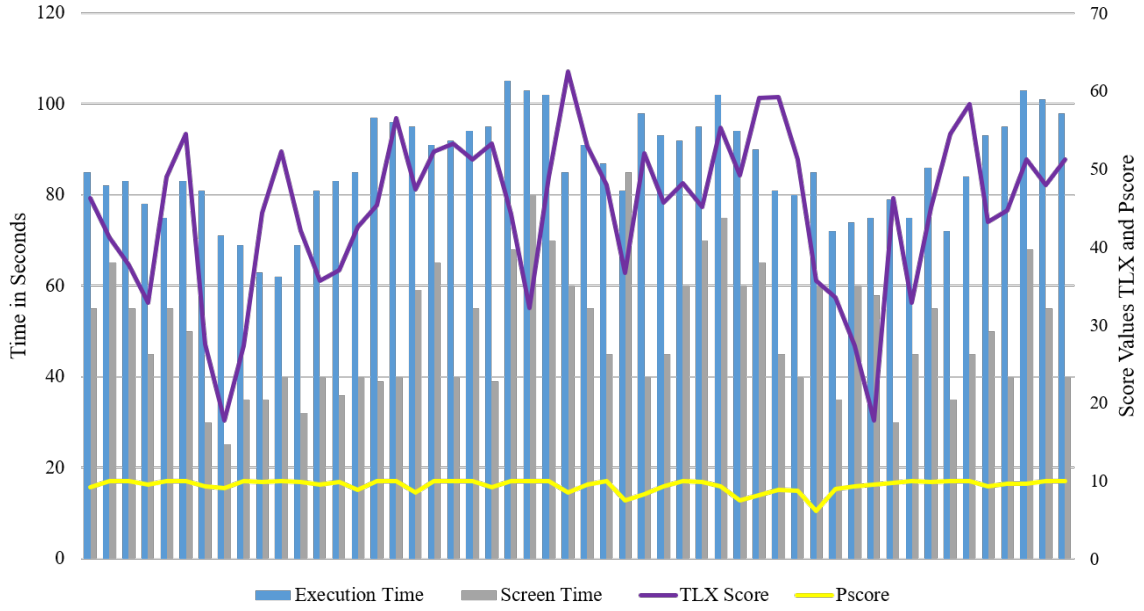

Figure 12. Screen time of all participants in comparison to execution time, TLX Score, and P-score of Trial 2

The two case studies enables us to draw inferences from several different perspectives. When it goes back to the research questions established in section 1.2, for research questions 1 we can use

 doi:10.20944/preprints202011.0352.v1

common metrics identified in [4] to measure performance of an HMT and its team members. For research question 2, in a controlled environment we are able to use VBSA which is developed in [8] as a team mate to form a successful HMT. Based on results we observed in this section, HIL can be used to study the performance of HMTs which effectively answering research questions 3. Based on the data analyzed in this section we are going to proposed guidelines to create standardized and well tested HMTs for practical application. These proposed guidelines still need to tested through rigorous testing and validation before accepting them as norms.

5. HMT Design and Standardization Guidelines

With insights gathered from the above two HIL simulation studies, an attempt has been made to provide guidelines for HMT architecting and implementation. This provides sufficient critical feedback that is required to develop the next key HMT guidelines for benchmarking and standardization. In this section, we present VBSA design guidelines, HMT architecting guidelines, and benchmarking guidelines. This section concludes with a summary of achievements of this work, application areas, and further discussion regarding possible future work directions in this area. These guidelines are formed based on statistical analysis and researchers observations from all our research publications [4,5,8,12,13,45–48] by the team to creating formulations to measure selected metrics.

In this section, we formulate guidelines that can be used to design a basic VBSA for critical applications by future researchers. Furthermore, we also formulate guidelines on how to develop an HMT using VBSA, as well as three mandatory steps that needed to be involved in constructing and using HMT team structures in real-world applications. However, these rules are still in development and testing stage, these guidelines need to go through rigorous experimental validation before accepting them as guidelines.

5.1. HMT Architecting and Implementation Guidelines

In general, the framing rules and roles of HMT components demands a well- defined architecture. Although HMT architecting can be done in numerous ways with variation in degrees of abstraction to retain the quality of HMT architectures, the following guidelines could be useful in constructing a VBSA that in turn could be used to develop HMTs.

5.1.1. Synthetic Agents Development Guidelines:

1. Through VBSA development and testing, we observe that structured language provides a significant burden of remembrance on humans in the team during interaction. Therefore, regarding natural language processing (NLP), strictly avoid using single structure language (i.e. level 1 NLP), although level 2 NLP can be allowed as a minimum voice interface in controlled environments. However, Level 3 NLP is recommended in all real world applications of VBSA.
2. Response time of VBSA for any action should be less than **300ms** in order to be applicable in real-time operations that include processing time in cloud services, communication, and conversions times. The minimum threshold response time should not be more than two seconds, as observed in the experimentation. In cases of more than two seconds, team members may get disengaged, or the VBSA may delay human natural operation capability.
3. In real world critical operation situations, neglect tolerance of VBSA should be **equal to task time** for optimum performance. As observed in the developed VBSA, the neglect tolerance is **36 seconds** maximum. It is currently sufficient for controlled environment operation and can be concluded as a neglect tolerance threshold value for a VBSA.
4. For critical tasks, cloud and probable delays should be calculated and included in a VBSA's normal response time to determine the total actual response time in which the VBSA takes task execution. Cloud delay will change based on service provider and place, so both need to be taken into consideration.

 doi:10.20944/preprints202011.0352.v1

5. Avoid false negatives and false positives in VBSA recognition in order to avoid human isolation of technology in real-world applications.

5.1.2. HMT Formation Guidelines:

1. Identifying clear-cut individual roles of both human and machine is the basic and primary element of HMT operations.
2. The technology incorporated in VBSA must also have room for customization to tailor meet any specific application scenario.
3. To qualify the term HMT, there should be at least one human and a SA. Therefore, identify the team combinations, i.e. number of human team members and SAs collaborating as a team.
4. Proper VBSA technology training should be provided to the human team members before proceeding to engage in a team task along with the VBSA.
5. Establish team operational rules to meet mission objectives without any occupational hazards to any of the team members.
6. A team training period is highly recommended to acclimatize team members on conduct of team operations and minimize interaction effort. This directly translates into improved team performance in real task accomplishment.
7. Test the task scenario with HMT to understand process flow deficiencies that exist, if any.
8. Validation trials: Performance measures developed for validation trials meant for HMT should have an ability to accommodate for any dynamic changes in work methods and performance objectives set by the human operator as the task load increases.

5.2. HMT Benchmarking and Standardization Guidelines

5.2.1. Minimum Requirements to Benchmark HMT:

1. Recognize all the parameters that are related to task success, and make necessary efforts to make sure the human-machine team acknowledge them during task execution.
2. Identify at least one method to measure the outcome of the HMT so that measures can be taken to address the anomalies in HMT task performance.
3. At a minimum, measure the accuracy of the results in order to determine error tolerance and define a threshold. Any order of error magnitude below the threshold value should not compromise task execution.
4. Try to identify any negative effects of interaction; otherwise, those unnoticed could lead to a cumulative workload burden on both human and machine.
5. At least one interaction metric is necessary to evaluate the interface management effect on the HMT task performance.
6. Identify minimum performance parameters that are required by an HMT to avoid task failure.

5.2.2. Performance Benchmarking Requirements:

1. Identify as many common parameters as possible to understand the in depth characteristics of an HMT.
2. Measure the HMT parameters using a combination of objective and subjective methods.
3. Measure performance through combinations of the overall process.
4. Study the effects of machine on human to establish clear weightage.
5. Machine situational awareness needs to be determined and then included in weightage.

6. Discussion and Conclusion

In essence, this work presents three significant contributions to the field of HMS research - design and development of task-oriented SA (i.e. VBSA) for HMT, development of methods in selecting the metrics for the purpose of benchmarking the studies of HMT, and demonstration of capability

 doi:10.20944/preprints202011.0352.v1

and usability of the VBSA to simulate HMT and performance measures. Several tasks that were accomplished to meet aforementioned contributions has been summarized below.

Section 2 illustrated the methodology for selecting common metrics to conduct performance studies of HMT metrics. As part of this section, a matrix of metrics was constructed, along with the realization of such metrics by arranging metrics according to their classifications and relations to HMT. The propose of such construction was to help the users with an easy access of the research material with metadata at one place. For the process of gathering metrics, we conducted a study of 185 published pieces of literature. The meta-analysis was done based on metric metadata as a measuring method. The analysis of metrics also includes the metric relations with HMT components, such as human, machine, and team, where a relation is established between metric and HMT, based on measurement types. As part of metric meta-analyses, we also identified inter-metric relations in which two metrics that may be using the same parameters could be used by researchers to carefully select one to measure performance while simultaneously avoiding multiple values during a single measurement. Moreover, in this section, a process was established to select common metrics to measure a set of HMT in which the formation of selection criteria proposed was based on two levels. Level 1 included HMT application scenarios, while Level 2 included selection criteria, metric metadata, and values. Through this process, one of the sets of common metrics proposed for the application scenarios contains ten common metrics in three components of HMT, presented as the results of Section 2. Readers who are interested in the selection process analysis as well as the detailed results and discussion thereof are encouraged to refer to [4].

Although most of the personal assistants commercially available today are voice-based, there are very limited work focused on SA work environments with architectures for emergency services. Moreover, full cloud-based VBSA development and overall performance analyses were never accomplished until we addressed it in our work. The detailed architecture design and development of the system in our work was aimed at system application requirements alongside the feedback of the users of the VBSA. The main motivation behind this work was to help users with minimal technical knowledge in this area to develop new scenarios and use them in different applications. Today, most of the technology that uses NLP systems is well developed, as it is critical in developing tools that can be used in a cutting-edge system, such as full-scale cloud architecture developed for VBSA. One of the limitations that was identified during the designing of the VBSA was the presence of vulnerabilities in the designed system, and those vulnerabilities have been analyzed and presented in Section 3. However, there are few facts that still remain unknown. Most notably, technology is changing at a rate like never before, which means that to get adapted and integrated changes into day-to-day systems, any single component change needs to be carefully evaluated in order to ensure safe operation and optimal performance of the system, as any mistakes can affect both human life and billions of dollars invested in the infrastructure. The development of a detailed cloud-based VBSA was the first achievement of this work and was published as [8].

Finally, HIL simulation of HMT was discussed in Section 4. Part of this section proposed an overview of HMT architectures and generalization of building blocks of the HMT. It also presented the supporting HMT architectures, published in [4]. In our work, the user studies or human-in-loop studies, was done considering two cases. In Case 1 of the study, the effects of a VBSA on HMT and on a human teammate was studied through performance comparison. The performance here was measured using user and scaling experiments by normalizing the scores obtained to a known measurement system. The discussion of the results presented, in this case, is already published in [8]. Case 2 of the study involved the use of VBSA in real-world scenarios and its effects, based on user expertise, language lexicon, implementation time, implementation errors, cognitive load, and workload using subjective and objective measurement techniques. Finally, as part of HIL simulations, detailed guidelines were presented on how to develop and construct new HMTs as well as how to standardize these HMTs in order to use them in critical environments. The guidelines suggest that VBSA is capable of becoming a teammate in critical situations and can be used to successfully simulate

 doi:10.20944/preprints202011.0352.v1

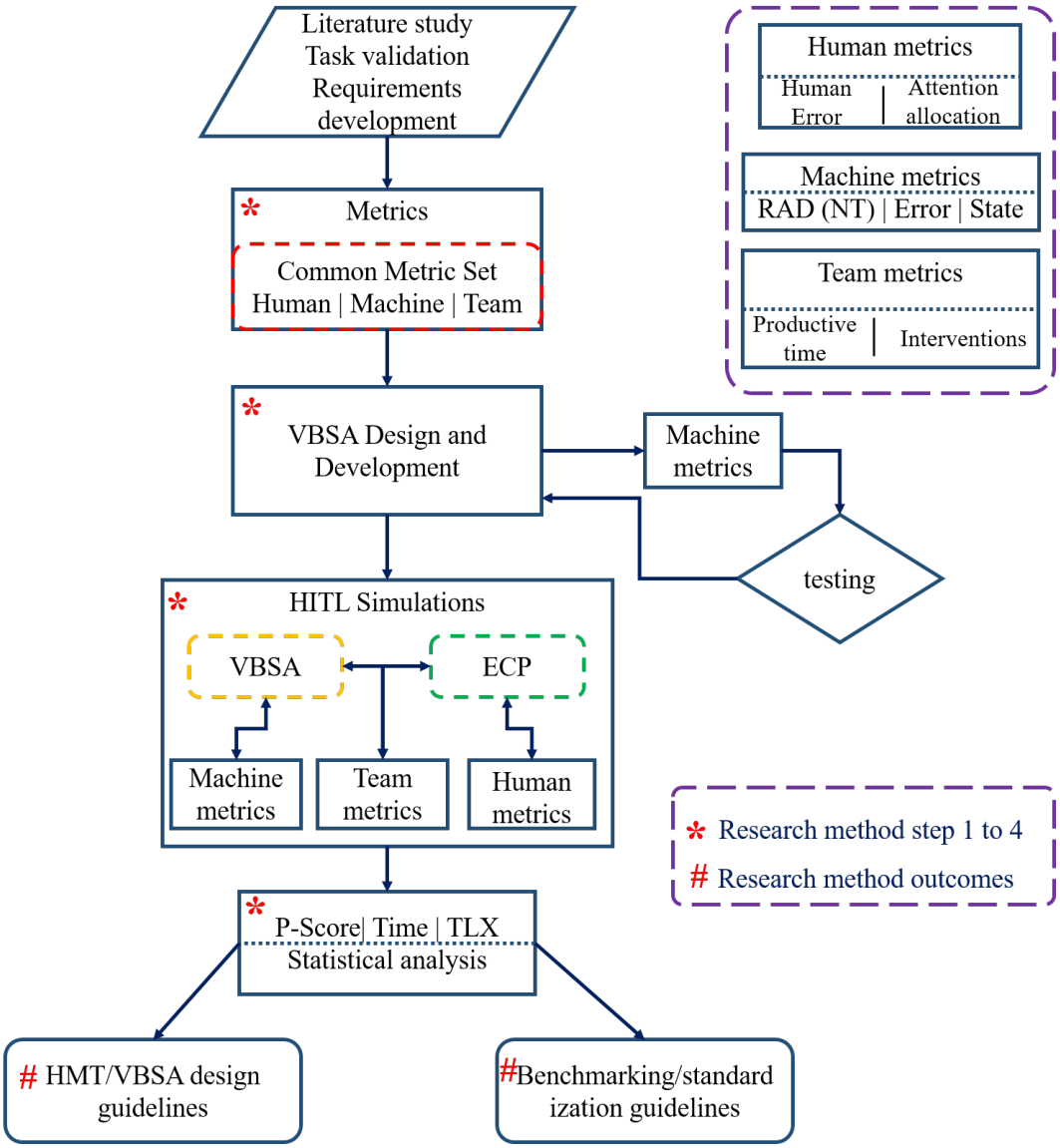

Figure 13. Hypothesized and validated HMT standardization method (reproduced from [5])

 doi:10.20944/preprints202011.0352.v1

HMTs to study real-world applications in a controlled environment. In addition, VBSA can be scaled to a large user base with minimal effort, which enables large teams to perform in parallel for multi-user simulations. Although a scenario with a maximum of 2 concurrent users was tested in our work, we hypothesize that VBSA will perform significantly well for a reasonably larger number of simultaneous users.

6.1. Future Work

As a future work in SA development, many different interfacing techniques can be integrated with the currently available cloud framework. However, each interaction would be different based on application scenarios. It is also worth considering human teammate limitations, such as eye-tracking or gesture interfaces, that can be used in environments in which safe voice interaction cannot be made. As part of VBSA enhancement, there should be more work on techniques to stabilize cloud response time, as during some days of operation, the response time of the cloud is too slow to be deemed as real-time, and in some situations, there is also a need for reduction of false negatives in NLP operation. Overall, providing better results and conclusions also requires improving the number of participants in the user study. Moreover, more metrics in data collection can be implemented in performance evaluation. Finally, as part of HMT standardization techniques, there is a substantial need of a study on the team training phase of HMTs, as many knowledge gaps still lie in this area and need to be filled.

Author Contributions: Conceptualization, Praveen Damacharla; Data curation, Parashar Dhakal and Jennie Gallimore; Formal analysis, Praveen Damacharla, Parashar Dhakal, Jyothi Priyanka Bandreddi, Jennie Gallimore and Colin Elkin; Funding acquisition, Ahmad Javaid and Vijay Devabhaktuni; Investigation, Praveen Damacharla, Ahmad Javaid and Vijay Devabhaktuni; Methodology, Praveen Damacharla, Ahmad Javaid and Vijay Devabhaktuni; Project administration, Ahmad Javaid; Resources, Ahmad Javaid, Colin Elkin and Vijay Devabhaktuni; Software, Parashar Dhakal; Supervision, Praveen Damacharla and Vijay Devabhaktuni; Validation, Jennie Gallimore; Writing – original draft, Praveen Damacharla; Writing – review & editing, Parashar Dhakal, Jyothi Priyanka Bandreddi, and Colin Elkin.

Funding: This study was funded by Round 1 Project Award "Improving Healthcare Training and Decision Making Through LVC" from the Ohio Federal Research Jobs Commission (OFMJC) through Ohio Federal Research Network (OFRN).

Acknowledgments: Authors appreciate support of the Paul A. Hotmer Family CSTAR (Cybersecurity and Teaming Research) Lab and EECS (Electrical Engineering and Computer Science) Department at the University of Toledo. Authors would like to thanks staff and management of Inter-professional Immersive Simulation Center (IISC) at The University of Toledo College of Medicine and Life Sciences for providing facilities and operational support. Authors also like to thank Ms. Dhwani Mehta, and Mr. Elliott Oberneder, for support in experimentation.

Conflicts of Interest: Authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript widely:

AI	Artificial Intelligence
ANOVA	the Two Way Analysis of Variance
AR	Augmented Reality
AVS	Alexa Voice Services
AWS	Amazon Web Services
ECP	Emergency Care Provider
FN	False Negative
FP	False Positive
HCI	Human-Computer Interface
HIL	Human-in-the-Loop
HMS	Human-Machine Systems
HMT	Human-Machine Teaming
IA	Intelligent Agents (or) Assistant
IRT	Item Response Theory
MD	Mental Demand
NLP	Natural Language Processing
PD	Physical Demand
P-Score	Performance Score
P-Value	Probability Value
SA	Synthetic Assistant (or) Agent
TD	Temporal Demand
TLX	Task Load Index
TN	True Negative
TP	True Positive
UML	Unified Modeling Language
VBSA	Voice Based Synthetic Assistant

References

1. Donovan, F. U.S. Military Employs AI, AR to Boost Medical Modeling, Simulation. https://hitinfrastructure.com/news/u.s.-military-employs-ai-ar-to-boost-medical-modeling-simulation/, 2019. [Online; accessed 17-May-2019].
2. Hudson, D.; Cohen, M. Use of intelligent agents in the diagnosis of cardiac disorders. 2002, Vol. 29, pp. 633 – 636. doi:10.1109/CIC.2002.1166852.
3. Ma, Y.; Chowdhury, M.; Sadek, A.; Jeihani, M. Real-Time Highway Traffic Condition Assessment Framework Using Vehicle–Infrastructure Integration (VII) With Artificial Intelligence (AI). *IEEE Transactions on Intelligent Transportation Systems* **2009**, *10*, 615–627.
4. Damacharla, P.; Javaid, A.Y.; Gallimore, J.J.; Devabhaktuni, V.K. Common metrics to benchmark human-machine teams (HMT): A review. *IEEE Access* **2018**, *6*, 38637–38655.
5. Damacharla, P.L.V.N. Simulation Studies and Benchmarking of Synthetic Voice Assistant Based Human-Machine Teams (HMT). Thesis, 2018.
6. Juliani, A.; Berges, V.P.; Vckay, E.; Gao, Y.; Henry, H.; Mattar, M.; Lange, D. Unity: A general platform for intelligent agents. *arXiv preprint arXiv:1809.02627* **2018**.
7. Gonçalves, E.; Araujo, J.; Castro, J. IStar4RationalAgents: Modeling requirements of multi-agent systems with rational agents. International Conference on Conceptual Modeling. Springer, 2019, pp. 558–566.
8. Damacharla, P.; Dhakal, P.; Stumbo, S.; Javaid, A.Y.; Ganapathy, S.; Malek, D.A.; Hodge, D.C.; Devabhaktuni, V. Effects of voice-based synthetic assistant on performance of emergency care provider in training. *International Journal of Artificial Intelligence in Education* **2019**, *29*, 122–143.
9. West, M.; Kraut, R.; Ei Chew, H. I'd blush if I could: closing gender divides in digital skills through education **2019**.
10. McNeese, N.J.; Demir, M.; Cooke, N.J.; Myers, C. Teaming With a Synthetic Teammate: Insights into Human-Autonomy Teaming. *Human Factors* **2018**, *60*, 262–273. PMID: 29185818, doi:10.1177/0018720817743223.
11. Showalter, D.E. More than Nuts and Bolts: Technology and the German Army, 1870–1945. *The Historian* **2002**, *65*, 123–143.
12. Damacharla, P.; Junuthula, R.R.; Javaid, A.Y.; Devabhaktuni, V.K. Autonomous ground vehicle error prediction modeling to facilitate human-machine cooperation. International Conference on Applied Human Factors and Ergonomics. Springer, 2018, pp. 36–45.
13. Damacharla, P.; Javaid, A.Y.; Devabhaktuni, V.K. Human error prediction using eye tracking to improvise team cohesion in human-machine teams. International Conference on Applied Human Factors and Ergonomics. Springer, 2018, pp. 47–57.
14. Licklider, J.C. Man-computer symbiosis. *Human Factors in Electronics, IRE Transactions on* **1960**, pp. 4–11.
15. Goertz, R.C. Manipulators used for handling radioactive materials. *Human factors in technology* **1963**, pp. 425–443.
16. Cipriani, C.; Zaccone, F.; Micera, S.; Carrozza, M.C. On the shared control of an EMG-controlled prosthetic hand: analysis of user–prosthesis interaction. *Robotics, IEEE Transactions on* **2008**, *24*, 170–184.
17. Griffiths, P.; Gillespie, R.B. Shared control between human and machine: Haptic display of automation during manual control of vehicle heading. Proceedings of 12th International Symposium on Haptic Interfaces for Virtual Environment and Teleoperator Systems (HAPTICS'04). IEEE, pp. 358–366.
18. Hoekenga, B.C. Mind over machine : what Deep Blue taught us about chess, artificial intelligence, and the human spirit. S m in science writing, 2007.
19. Sullivan, M.; Masters, T.; Anderson, P.; Bennett, J.; Bonner, M.; Hassinger, K.; Porter, M.; Suding, M.; Volk, A. F-35 Joint Strike Fighter: Assessment Needed to Address Affordability Challenges. Report, DTIC Document, 2015.
20. Hicks, J.S.; Durbin, D.B. An Investigation of Multiple Unmanned Aircraft Systems Control from the Cockpit of an AH-64 Apache Helicopter. Report ARL-TR-7151, DTIC Document, 2014.
21. Klaproth, O.W.; Halbrügge, M.; Krol, L.R.; Vernaleken, C.; Zander, T.O.; Russwinkel, N. A Neuroadaptive Cognitive Model for Dealing With Uncertainty in Tracing Pilots' Cognitive State. *Topics in Cognitive Science* **2020**, *12*, 1012–1029.

22. Novitzky, M.; Robinette, P.; Gleason, D.K.; Benjamin, M.R. A platform for studying human-machine teaming on the water with physiological sensors. Workshop on Human-Centered Robotics: Interaction, Physiological Integration and Autonomy at RSS 2017, 2017.
23. Le Vie, L.R.; Last, M.C.; Barrows, B.; Allen, B.D. Towards Informing an Intuitive Mission Planning Interface for Autonomous Multi-Asset Teams via Image Descriptions. 2018 Aviation Technology, Integration, and Operations Conference, 2018, p. 4013.
24. O'Neill, T.; McNeese, N.; Barron, A.; Schelble, B. Human–Autonomy Teaming: A Review and Analysis of the Empirical Literature. *Human Factors* **2020**, p. 0018720820960865.
25. Demir, M.; McNeese, N.J.; Johnson, C.; Gorman, J.C.; Grimm, D.; Cooke, N.J. Effective team interaction for adaptive training and situation awareness in human-autonomy teaming. 2019 IEEE Conference on Cognitive and Computational Aspects of Situation Management (CogSIMA). IEEE, 2019, pp. 122–126.
26. Dubey, A.; Abhinav, K.; Jain, S.; Arora, V.; Puttaveerana, A. HACO: A Framework for Developing Human-AI Teaming. Proceedings of the 13th Innovations in Software Engineering Conference on Formerly known as India Software Engineering Conference, 2020, pp. 1–9.
27. Fiore, S.M.; Wiltshire, T.J. Technology as teammate: Examining the role of external cognition in support of team cognitive processes. *Frontiers in psychology* **2016**, *7*, 1531.
28. Grimm, D.A.; Demir, M.; Gorman, J.C.; Cooke, N.J. Team Situation Awareness in Human-Autonomy Teaming: A Systems Level Approach. Proceedings of the Human Factors and Ergonomics Society Annual Meeting. SAGE Publications Sage CA: Los Angeles, CA, 2018, Vol. 62, pp. 149–149.
29. Wohleber, R.W.; Stowers, K.; Chen, J.Y.; Barnes, M. Effects of agent transparency and communication framing on human-agent teaming. 2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC). IEEE, 2017, pp. 3427–3432.
30. Demir, M.; McNeese, N.J.; Cooke, N.J. Team synchrony in human-autonomy teaming. International Conference on Applied Human Factors and Ergonomics. Springer, 2017, pp. 303–312.
31. Jay, S.R.; Brandt, S.L.; Lachter, J.; Matessa, M.; Sadler, G.; Battiste, H. Application of human-autonomy teaming (HAT) patterns to reduced crew operations (RCO). International Conference on Engineering Psychology and Cognitive Ergonomics. Springer, 2016, pp. 244–255.
32. Demir, M.; McNeese, N.J.; Cooke, N.J. Team situation awareness within the context of human-autonomy teaming. *Cognitive Systems Research* **2017**, *46*, 3–12.
33. Singh, H.V.P.; Mahmoud, Q.H. ViDAQ: A Framework for Monitoring Human Machine Interfaces. 2017 IEEE 20th International Symposium on Real-Time Distributed Computing (ISORC), 2017, pp. 141–149. doi:10.1109/ISORC.2017.25.
34. Ren, J.; Vlachos, T.; Argyriou, V. Immersive and perceptual human-computer interaction using computer vision techniques. Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on. IEEE, 2010, pp. 66–72.
35. Echeverria, G.; Lassabe, N.; Degroote, A.; Lemaignan, S. Modular open robots simulation engine: Morse. Robotics and Automation (ICRA), 2011 IEEE International Conference on. Citeseer, 2011, pp. 46–51.
36. Nikolaidis, S.; Shah, J. Human-robot cross-training: computational formulation, modeling and evaluation of a human team training strategy. Proceedings of the 8th ACM/IEEE international conference on Human-robot interaction. IEEE Press, 2013, pp. 33–40.
37. Cummings, M.L. The need for command and control instant message adaptive interfaces: Lessons learned from Tactical Tomahawk human-in-the-loop simulations. *CyberPsychology & Behavior* **2004**, *7*, 653–661.
38. Meserole, J.S.; Moore, J.W. What is system wide information management (SWIM)? 25th Digital Avionics Systems Conference, 2006 IEEE/AIAA. IEEE, 2006, pp. 1–8.
39. Von Campenhausen, C.; Petrisch, G. The benchmarking matrix. *Managerial Auditing Journal* **2004**, *19*, 172–179.
40. Orr, D. A Report of the working Group on Evaluating Team performance. *Interagency Advisory Group [IAG] Committee on Performance Management and Recognition. Gary McLean, FCC (202)* **1993**, pp. 632–7541.
41. Strybel, T.Z.; Keeler, J.; Mattoon, N.; Alvarez, A.; Barakezyan, V.; Barraza, E.; Park, J.; Vu, K.P.L.; Battiste, V. Measuring the Effectiveness of Human Autonomy Teaming. International Conference on Applied Human Factors and Ergonomics. Springer, 2017, pp. 23–33.

42. V P Singh, H.; Mahmoud, Q.H. NLP-Based Approach for Predicting HMI State Sequences Towards Monitoring Operator Situational Awareness. *Sensors (Basel, Switzerland)* **2020**, *20*, 3228. doi:10.3390/s20113228.
43. Callaway, D.W.; Smith, E.R.; Cain, J.; Shapiro, G.; Burnett, W.T.; McKay, S.D.; Mabry, R. Tactical emergency casualty care (TECC): guidelines for the provision of prehospital trauma care in high threat environments. *J spec oper med* **2011**, *11*, 104–122.
44. Savage, L.E.; Forestier, M.C.; Withers, L.N.; Tien, C.H.; Pannell, C.D. Tactical combat casualty care in the Canadian Forces: lessons learned from the Afghan war. *Canadian Journal of Surgery* **2011**, *54*, S118.
45. Dhakal, P.; Damacharla, P.; Javaid, A.Y.; Devabhaktuni, V. Detection and Identification of Background Sounds to Improvise Voice Interface in Critical Environments. 2018 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT). IEEE, 2018, pp. 078–083.
46. Dhakal, P.; Damacharla, P.; Javaid, A.Y.; Devabhaktuni, V. A near real-time automatic speaker recognition architecture for voice-based user interface. *Machine Learning and Knowledge Extraction* **2019**, *1*, 504–520.
47. Dhakal, P. Novel Architectures for Human Voice and Environmental Sound Recognition using Machine Learning Algorithms. PhD thesis, University of Toledo, 2018.
48. Damacharla, P.; Mehta, D.; Javaid, A.Y.; Devabhaktuni, V. Study on State-of-the-art Cloud Systems Integration Capabilities with Autonomous Ground Vehicles **2018**. pp. 1–5.
49. Nikolaidis, S.; Ramakrishnan, R.; Gu, K.; Shah, J. Efficient model learning from joint-action demonstrations for human-robot collaborative tasks. Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction. ACM, 2015, pp. 189–196.
50. Gombolay, M.C.; Gutierrez, R.A.; Clarke, S.G.; Sturla, G.F.; Shah, J.A. Decision-making authority, team efficiency and human worker satisfaction in mixed human–robot teams. *Autonomous Robots* **2015**, *39*, 293–312.
51. Bruemmer, D.J.; Walton, M.C. Collaborative tools for mixed teams of humans and robots. Technical report, IDAHO NATIONAL ENGINEERING AND ENVIRONMENTAL LAB IDAHO FALLS, 2003.
52. Harriott, C.E.; Adams, J.A. Modeling Human Performance for Human–Robot Systems. *Reviews of Human Factors and Ergonomics* **2013**, *9*, 94–130.
53. Manning, M.D.; Harriott, C.E.; Hayes, S.T.; Adams, J.A.; Seiffert, A.E. Heuristic Evaluation of Swarm Metrics' Effectiveness. Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction Extended Abstracts. ACM, 2015, pp. 17–18.
54. Burke, J.L.; Murphy, R.R.; Rogers, E.; Lumelsky, V.J.; Scholtz, J. Final report for the DARPA/NSF interdisciplinary study on human-robot interaction. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on* **2004**, *34*, 103–112.
55. Harriott, C.E.; Seiffert, A.E.; Hayes, S.T.; Adams, J.A. Biologically-inspired human-swarm interaction metrics. Proceedings of the Human Factors and Ergonomics Society Annual Meeting. SAGE Publications, 2014, Vol. 58, pp. 1471–1475.
56. Ball, J.; Myers, C.; Heiberg, A.; Cooke, N.J.; Matessa, M.; Freiman, M.; Rodgers, S. The synthetic teammate project. *Computational and Mathematical Organization Theory* **2010**, *16*, 271–299.
57. Cacciabue, P.C. Elements of human-machine systems. In *Guide to Applying Human Factors Methods*; Springer, 2004; pp. 9–47.
58. Gombolay, M.C.; Wilcox, R.J.; Shah, J.A. Fast scheduling of robot teams performing tasks with temporospatial constraints. *IEEE Transactions on Robotics* **2018**, *34*, 220–239.
59. Kim, J.; Shah, J.A. Improving team's consistency of understanding in meetings. *IEEE Transactions on Human-Machine Systems* **2016**, *46*, 625–637.
60. Sen, S.D.; Adams, J.A. Real-Time Optimal Selection of Multirobot Coalition Formation Algorithms Using Conceptual Clustering. Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence, 2015.
61. Wiltshire, T.J.; Fiore, S.M. Social Cognitive and Affective Neuroscience in Human–Machine Systems: A Roadmap for Improving Training, Human–Robot Interaction, and Team Performance. *Human-Machine Systems, IEEE Transactions on* **2014**, *44*, 779–787.
62. Murphy, R.R.; Schreckenghost, D. Survey of metrics for human-robot interaction. Human-Robot Interaction (HRI), 2013 8th ACM/IEEE International Conference on. IEEE, 2013, pp. 197–198.