Artificial Intelligence in Education 2022
Durham, England, UK
July 31st, 2022

Virtual Workshop Proceedings: Advances and Opportunities in Team Tutoring

Workshop Chair:
Anne M. Sinatra, Ph.D.
US Army Combat Capabilities Development Command (DEVCOM) Soldier Center

Workshop Co-Chair:
Benjamin Goldberg, Ph.D.
US Army Combat Capabilities Development Command (DEVCOM) Soldier Center

Workshop Committee:

Keith Brawner, Ph.D.
US Army Combat Capabilities Development Command (DEVCOM) Soldier Center

Gregory Goodwin, Ph.D.
US Army Combat Capabilities Development Command (DEVCOM) Soldier Center

Proceedings Edited by Anne M. Sinatra and Benjamin Goldberg

The views expressed herein are those of the authors and do not necessarily reflect the views of the U.S. Army Combat Capabilities Development Command – Soldier Center

Use of trade names or names of commercial sources is for information only and does not imply endorsement by the U.S. Army Combat Capabilities Development Command – Soldier Center.

This publication is intended to provide accurate information regarding the subject matter addressed herein. The information in this publication is subject to change at any time without notice. The U.S. Army Combat Capabilities Development Command – Soldier Center, nor the authors of the publication, makes any guarantees or warranties concerning the information contained herein.

Printed in the United States of America
First Printing, January 2023

U.S. Army Combat Capabilities Development Command
Soldier Center
Orlando, Florida

Preface

This virtual workshop was conducted in conjunction with the 22nd Artificial Intelligence in Education Conference (AIED). The AIED 2022 conference was a hybrid conference being held both in person at Durham University in England, and online. This workshop was an entirely online virtual workshop that occurred as part of the conference. This workshop focused on different applications of team tutoring within intelligent tutoring systems (ITSs). The workshop included examples of approaches that have been used to implement team tutoring, and approaches that have been used to overcome the challenges of team tutoring. The workshop included a discussion of commonalities in the presented work, and the next steps forward for team tutoring research. Work presented at the workshop is included in the following proceedings papers.

July 2022
Anne M. Sinatra and Benjamin Goldberg

Acknowledgement

The statements and opinions expressed in this publication do not necessarily reflect the position or the policy of the United States Government, and no official endorsement should be inferred.

Table of Contents

Toward Competency-Driven Team Performance Measurement and Coaching in GIFT

Randall Spain[1], Andy Smith[1], Jonathan Rowe[1], Benjamin Goldberg[2], and James Lester[1]

[1] Center for Educational Informatics, North Carolina State University, Raleigh NC 27695, USA
[2] {rdspain, smith, jprowe, lester}@ncsu.edu
[2] U.S. Army DEVCOM – Soldier Center, Orlando FL 32826, USA
benjamin.s.goldberg.civ@mail.mil

Abstract. A critical step toward the effective assessment of team performance is identifying the set of team competencies that should be measured during training to guide team performance diagnosis. This work describes our application of a model of team competencies in an assessment framework to drive automated team-based coaching capabilities in a crew gunnery training course we are creating using the Generalized Intelligent Framework for Tutoring. We present an assessment model that identifies crew coordination competencies at the individual and team level and discuss team process and team outcome measures we are using to measure and diagnose crew performance during training.

Keywords: Team performance assessment, teamwork, teamwork competencies.

1 Introduction

Defining team competencies is a principal step in developing team performance measurement. Competencies reflect measurable patterns of knowledge, skills, abilities (KSAs), behaviors and other characteristics that individuals need in order to perform work roles or job functions successfully. Team competencies reflect the collective KSAs team members must demonstrate to accomplish shared goals. Examples of team competencies include shared task models, knowledge of team role interaction patterns, shared situation awareness, and team cohesion [1]. To promote the development of teamwork skills, adaptive instructional systems (AISs) for teams must include assessment models that map teamwork competencies to observable actions, conditions, or state information that reflect evidence of these KSAs. Defining these models can be a challenging task, as most team competencies are multifaceted and require a keen understanding about the interplay between the teamwork and taskwork required to support effective team performance [2, 3]. In addition, AISs have historically been designed to support individual assessment and learning. Extending assessment models and frameworks to support team training is a current area of active research.

In this paper we discuss how we have adopted a model of team competencies developed by Cannon-Bowers et al. [2] to inform the creation of an assessment framework to drive automated team-based coaching capabilities in a course we are creating using the Generalized Intelligent Framework for Tutoring (GIFT). The team-based coaching framework is being designed to provide automated coaching to gunnery crews performing crew gunnery training exercises in Virtual Battlespace 3 (VBS3), a simulation-based training environment used by the U.S. Army. The framework distinguishes different types of team competencies and has implications regarding whether assessments should be modeled at the individual or team level.

In addition to describing our use of Cannon-Bowers et al.'s model to guide the development of a team assessment and feedback framework, we discuss how the distinctions among team performance outcomes and processes have implications for defining reward functions in reinforcement learning-driven pedagogical planning frameworks in AISs for teams. We conclude with an overview of our current assessment model and by discussing upcoming research activities that aim to evaluate the effectiveness of different forms of coaching actions on team training performance.

2 Related Work

A critical step towards the development of effective team training scenarios in AISs is identifying the team competencies that should be assessed during training to facilitate team performance diagnosis and remediation [3, 4].

In this work, team performance diagnosis refers to the process of accurately linking, measuring, and interpreting the behaviors and changes that lead to effective and ineffective performance during training. Teams have characteristics and engage in behaviors that distinguish them from individuals. These characteristics and behaviors, often referred to as team competencies, mediate and moderate team performance and team effectiveness.

At last year's AIED team tutoring workshop, we discussed ongoing research using GIFT to investigate reinforcement learning (RL) based coaching policies for promoting team performance in the domain of crew gunnery training [5]. We described our approach towards collecting a corpus of multimodal training data including video, communication, and simulation-trace data from the U.S. Army gunnery crews who are completing simulation-based training exercises to prepare for crew gunnery qualification and how we planned to use the dataset to induce data-driven coaching policies for promoting individual and crew gunnery performance.

A key objective of our team's work over the past year has been to develop and integrate a crew gunnery assessment model into our GIFT-based course that can guide the assessment of crew performance and diagnose deficiencies in team performance. Our goal has been to measure team processes and outcomes and to assess performance at the team and individual levels. In this work, we adapt the definition consistent with previous team science research in that the term process refers to "the collection of activities, strategies, and behaviors employed in [the] task accomplished" [3], whereas team outcomes refer to the outputs of team performance [6].

There are several reasons for including team process measures in an assessment model in addition to team outcome measures. First, process-oriented measures allow evaluators (e.g., human observers and AIS-based systems) to gather evidence regarding the actions and behaviors a team engaged in to achieve a specific outcome. These measures allow for a granular diagnosis of flaws that may have led to a team's subpar performance. Focusing solely on outcome measures, such as task completion time or task completion accuracy, overlooks *how* team behaviors affected performance.

Second, flawed processes can occasionally result in successful performance. For instance, a team may reach the outcome standards set for "passing" training but, the process the team engaged in to reach these standards may not reflect effective teamwork skills. In these scenarios, flawed behaviors and team actions can be mistakenly reinforced [3]. These types of oversights have significant implications for devising data-driven coaching models that aim to provide coaching and feedback to improve team processes and team performance. If the model is developed by only considering team performance outcomes alone, then the model could ultimately reinforce flawed team processes which could result in negative training value. Therefore, AIS developers should consider assessing both team processes and team performance outcomes.

To guide the development of the team measurement model we have adopted a framework developed by Cannon-Bowers et al. [2] and expanded upon by Cannon-Bowers and Salas [3] that delineates different levels of team competencies. The first level includes individual-level KSAs required to perform job functions. The focus here is on taskwork concepts, such as task specific role responsibilities and procedural knowledge, that are needed at the individual level to support effective team performance. The second level includes team-oriented competencies that are held at the individual level. These competencies are transportable across teams (team-generic) and directly influence team performance regardless of the team member involved. Examples include communication skills, leadership skills, and attitudes towards teamwork. An example is information exchange, which is a dimension of team development, but is performed at the individual level and should thus be measured and remediated at the individual level. The third level includes team-oriented KSAs that are held at the team level and are specific to the task and team involved (context-driven competencies). Examples include skills such as shared situation awareness and attitudes such as team cohesion.

The distinction between these levels is important as it has implications for what is measured and at what level the team competencies are measured (e.g., is the skill measured at the individual or team level?). It also has implications for how feedback and coaching decisions are derived and delivered to teams to support the development of effective teamwork skills. For example, effective communication may be conceptualized as an important teamwork skill, but one that is held at the individual level, and therefore best measured (and trained and remediated) at the individual level [3]. On the other hand, collective efficacy (i.e., the team's belief that it can cope with task demands) is rooted at the team level and is team specific, and can be measured meaningfully only at the team level. Consequently, effective feedback and coaching to enhance collective efficacy may best be delivered at the team level as well.

In summary, identifying and conceptualizing team performance measures according to whether they reflect individual or team-level competencies and whether they are process- or outcome-oriented is a necessary step toward the development of competency-driven assessment approaches for team-training. In the next section, we discuss our work toward developing an assessment in GIFT to support crew gunnery training.

3 Team Performance Measurement for Crew Gunnery in GIFT

Team performance measures should be able to describe, evaluate, and diagnose team performance. This is a necessary prerequisite and a critical challenge for devising computational models that provide instructional support and coaching in AISs for teams.

In recent years, GIFT has become an important tool for investigating how these challenges can be addressed. GIFT is an open-source service-oriented framework for designing, developing, and evaluating AISs [7]. GIFT provides developers with a suite of authoring tools for rapidly creating adaptive courses to support individual and team training events in web-based and simulation-based training platforms. Recent enhancements to GIFT support team tutoring opportunities. Notable enhancements include the ability to model team structures in GIFT's assessment architecture, the addition of new condition classes and scenario adaptations to support automated assessment of team behaviors and team performance in simulation-based training environments, and the ability to deliver adaptive feedback and coaching messages at the individual and team level [8].

3.1 Crew Gunnery Training and Assessment

In our current line of work, we are investigating how data-driven models can be devised to provide crews with automated coaching and guided instruction to support the acquisition of crew coordination skills [9–11]. We have developed a crew gunnery course using GIFT that evaluates crew performance as they complete crew gunnery training tables in VBS3. VBS3 is an immersive virtual training environment that the U.S. Army uses to support individual and collective training. Specifically, we are using a set of VBS3 missions that emulate real-world training and qualification courses that U.S. Army gunnery crews complete in preparation for live-fire training and qualification (Figure 1). The virtual scenarios were developed at the Warrior Skills Training Center (WSTC) in Fort Hood, Texas and offer crews the opportunity to practice and rehearse crew coordination activities in a controlled environment.

The VBS3 crew gunnery scenario involves a series of six engagements. Each engagement requires gunnery crews to coordinate actions in order to carry out the direct fire engagement process as outlined in TC 3-20.31-4 (Direct Fire Engagement Process). A key component of the direct fire engagement process is the coordination of actions and behaviors among the vehicle commander, gunner, and driver. During an engagement, crew members coordinate actions and exchange information pertaining to potential threats in the environment. When a threat has been identified, crews engage in a fire command sequence, which is a well-defined protocol for communicating information and actions to facilitate a coordinated response to a threat. During the fire command process, the vehicle commander and gunner follow a prescriptive set of directives and spoken commands and acknowledgements for engaging identified targets.

Fig. 1. Crew Gunnery Training in VBS3

The current U.S. Army gunnery standards, i.e. TC 3-20.31 (Training and Qualification Crew) provides standards that guide the evaluation of crew performance. Crews earn points that range from 0 to 100 points for each engagement. The points are calculated using a rubric that incorporates a combination of the firing vehicle's position, target type, target movement, target range, and target neutralization time to determine the engagement score. Timely engagement and termination of targets is critical for earning high evaluation scores. Crews can also receive point deductions for violating the actions outlined in the fire command protocol and for committing safety violations. After each engagement the points are aggregated to determine if the engagement was "passed", which is achieved by scoring 70 or more points.

This point-based approach to evaluation reflects a training outcome measure. But as indicated in an earlier section of our discussion, including both outcome and process measures is critical for developing team performance measures to support effective diagnosis of crew performance and support remediation. To support a more granular analysis of crew performance and to better measure the processes that impact crew performance, we have developed a crew assessment model that includes a series of concepts and supporting subconcepts that align to the Observe, Detect, Engage, and Report phases of the direct engagement process. These concepts are further classified into engagement-specific activities that crew members must complete. To guide the development of our assessment model we reviewed the crew gunnery training and qualification curriculum (TC 3-20.31), reviewed relevant field manuals, interviewed former crew gunnery evaluators and Subject Matter Experts (SMEs), and reviewed a task analysis of the crew gunnery domain [12].

3.2 Aligning Crew Competencies to Team Performance Measurement Levels

As previously discussed, measuring performance at the individual level and at the team level is a recommended best practice for team performance measurement [3]. Table 1 provides a selected set of team competencies that we are modeling in our crew gunnery assessment model in GIFT. The table includes the course concept associated with the competency and a description of the guiding assessment question. Utilizing the framework proposed by Cannon-Bowers et al. [2] and further discussed by Cannon-Bowers and Salas [3], we have also distinguished individual-level competencies, team competencies held at the individual level, and team competencies held at the team level to guide our measurement model. The competencies are organized into two general phases of the direct fire engagement process: Observe and Engage.

During the Observe phase, crew members are required to search and scan the environment for potential threats. Crew members should engage in coordinated search behaviors so that sectors are fully and accurately scanned. To facilitate diagnosis of crew scan behavior, we are utilizing two condition classes in GIFT: assigned sector and detect object. The *assigned sector* condition class assesses whether team members are searching within their assigned sectors of responsibility [8]. The *detect object* condition class assesses how long it takes one or more team members to properly detect an object (i.e., potential threat) after the object has entered the team member's field of view [8].

These measures provide insights and a means for assessing team members' observation and scanning behaviors that can affect target engagement times. Another important teamwork skill we are assessing during this phase of the task is crew coordination. Specifically, we are examining inter-crew communication to identify if and how crews coordinate actions in order to establish scanning responsibilities during the direct fire engagement process. Assessing these behaviors is critical for diagnosing and remediating crew performance, because a breakdown in crew coordinated search could impact target engagement times and overall engagement scores.

In addition to assessing crew coordination during the Observe phase, we are measuring critical crew coordination behaviors that occur during the Engage phase of the direct fire engagement process. Once a threat has been identified, crews engage in a fire command sequence, which is a well-defined protocol for communicating information and actions to facilitate a coordinated response to a threat. Here we are interested in measuring crew communication and information exchange between team members. Because the crews are trained to follow a specific verbal protocol it is critical to assess whether crews are using correct terminology and engaging in closed loop communication. These three facets of crew communication – information exchange, phraseology and closed loop communication – can lead to communication breakdowns when not applied appropriately and ultimately impact shared cognition among crew members and crew performance.

Table 1. Crew gunnery team performance competencies and measurement levels.

Engagement Phase	Course Concept	Guiding Question	Team Competency	Measurement Level
Observe	Search assigned sector	Are crew members scanning their assigned sector?	Task-specific role responsibility	Individual level competency
Observe	Overlapping sectors	Did crew members coordinate scan techniques to ensure sector search (scanning) are overlapping?	Information Exchange	Team competency held at the individual level
Observe	Target identification	Did the crews accurately recognize threats in the environment?	Cue-strategy association	Individual-level competency
Engage	Accurate and complete fire command	Did the crew member exchange the right type of information to the right person during the engagement?	Information Exchange	Team-competency held at the individual level
Engage	Accurate and complete fire command	Was the fire command succinct and did it contain correct terminology?	Phraseology	Individual-level competency
Engage	Coordinated actions during the fire command	Did the crew commit any procedural violations during the engagement?	Team role-interaction patterns	Team-competency at the individual level

The current iteration of our crew assessment model utilizes an observed assessment condition class to facilitate the measurement of crew communication and coordination for the fire command sequence tasks. The observed assessment condition class allows a human observer to assess concepts defined in a GIFT-based training course using the GIFT GameMaster interface. This interface facilitates a "human in the loop" AIS interaction model for assessing performance and injecting scenario adaptations during collective simulation-based training events. GIFT GameMaster enables observer controllers to visualize unit progress through a live map view; monitor completed tasks, active tasks, and upcoming tasks; provide manual performance assessments; and inject scenario adaptations to shape unfolding simulation-based training scenarios at run-time [8]. GameMaster also enables instructors to provide pre-configured feedback messages and coaching, or tailor their own message, and select to whom feedback and coaching is delivered (e.g., an individual, a subset of the team, the entire team). A significant goal of our current research program is to replace this manual-assessment component with a natural language processing-based team communication analysis approach that can be used to automatically assess crew communication practices during the VBS3 training exercise [13].

In addition to course concepts outlined in Table 1 which reflect team process measures, we are also measuring task-specific team performance outcomes. Specifically, we are measuring (a) the amount of time in seconds it takes the crew to issue an alert that a target is in the environment and (b) the length of time in seconds it takes crews to terminate the target once the target is in the locked position. Together, these team process and team outcome measures are being used to investigate rewards that can be modeled in a data-driven framework to facilitate automated coaching [10].

An advantage of using the framework proposed by Cannon-Bowers et al. [2] to identify team training measurement levels is that it provides a theory-driven framework for formalizing where team performance should be measured and where remediation should be directed. We have found it to be especially helpful for developing an assessment framework that links course concepts with team process measures as well as feedback and coaching statements that can be delivered to crew members. In addition, it links team-generic and team-specific competencies to KSAs that can impact team performance. Team-generic competencies refer to competencies held by individual team members and can influence team performance regardless of the teammate involved. They are transportable across teams. Examples include communication skills, leadership skills, and attitudes towards teamwork. Team-specific competencies refer to the set of competencies that are required to function effectively on a specific team. These are content-specific behaviors and processes that need to be demonstrated to perform effectively within the team. Developers of AISs for team tutors can use these distinctions to identify competencies that are likely to be important for different types of team training objectives.

4 Conclusion and Future Direction

A critical step toward the effective assessment of team performance is identifying the set of team competencies that should be assessed during training to guide team performance diagnosis. In this paper we discuss how we are utilizing a competency framework developed in previous team science research to guide the development of a crew gunnery assessment model for an adaptive training course.

In the upcoming months, we will collect training interaction data utilizing the GIFT-based crew gunnery course we are developing and use this data to investigate the effectiveness of different forms of adaptive team-based feedback and coaching. We will utilize the crew gunnery assessment model to facilitate team performance measurement and evaluate their effectiveness for developing and revising data-driven team-based assessment models in AISs.

Acknowledgment
The research described herein has been sponsored by the U.S. Army Combat Capabilities Development Command, Soldier Center under cooperative agreement W912CG-19-2-0001. The statements and opinions expressed in this article do not necessarily reflect the position or the policy of the United States Government, and no official endorsement should be inferred.

References

1. Shuffler, M.L., Pavlas, D., Salas, E.: Teams in the military: A review and emerging challenges. In: The Oxford handbook of military psychology. pp. 282–310. Oxford University Press, New York, NY, US (2012). https://doi.org/10.1093/oxfordhb/9780195399325.013.0106.
2. Cannon-Bowers, J.A., Tannenbaum, S.I., Salas, E., Volpe, C.E.: Defining competencies and establishing team training requirements. In: Team effectiveness and decision making in organizations. pp. 333–380. Wiley (1995).
3. Cannon-Bowers, J., Salas, E.: A Framework for Developing Team Performance Measures in Training. Psychology Press (1997). https://doi.org/10.4324/9781410602053-10.
4. Goldberg, B., Owens, K., Gupton, K., Hellman, K.: Forging Competency and Proficiency through the Synthetic Training Environment with an Experiential Learning for Readiness Strategy. 15 (2021).
5. Spain, R.D., Rowe, J.P., Goldberg, B.S., Pokorny, R., Hoffman, M., Harrison, S., Lester, J.C.: Developing Adaptive Team Coaching in GIFT: A Data-Driven Approach. In: TTW@ AIED. pp. 10–16 (2021).
6. Delise, L.A., Allen Gorman, C., Brooks, A.M., Rentsch, J.R., Steele-Johnson, D.: The effects of team training on team outcomes: A meta-analysis. Perform. Improv. Q. 22, 53–80 (2010). https://doi.org/10.1002/piq.20068.
7. Sottilare, R.A., Brawner, K., Sinatra, A., Johnston, J.: An Updated Concept for a Generalized Intelligent Framework for Tutoring (GIFT). GIFT Users Symp. 19 (2017).
8. Hoffman, M., Goldberg, B., Brawner, K.: The GIFT Architecture and Features Update: 2022 Edition. In: Generalized Intelligent Framework for Tutoring (GIFT) Users Symposium (GIFTSym10). p. 11 (2021).
9. Rowe, J., Spain, R., Goldberg, B., Pokorny, R., Mott, B., Lester, J.: Toward Data-Driven Models of Team Feedback in Synthetic Training Environments with GIFT. In: Proceedings of the 8th Annual Generalized Intelligent Framework for Tutoring (GIFT) Users Symposium (GIFTSym8). p. 136. US Army Combat Capabilities Development Command–Soldier Center (2020).
10. Smith, A., Spain, R.D., Rowe, J., Goldberg, B., Lester, J.: Formalizing Adaptive Team Feedback in Synthetic Training Environments with Reinforcement Learning. In: Generalized Intelligent Framework for Tutoring (GIFT) Users Symposium (GIFTSym10). p. 117 (2022).
11. Spain, R., Rowe, J., Goldberg, B., Pokorny, R., Mott, B., Lester, J.: Automated coaching in synthetic training environments: Developing an adaptive team feedback framework. In: Proceedings of the ninth annual GIFT users symposium (GIFTSym9). pp. 187–199 (2021).
12. Morrison, J.E., Meade, G.A., Campbell, R.C.: Identification of Crew-and Platoon-Level Gunnery Subtasks: Objectives for a Threat-Based Training Program. HUMAN RESOURCES RESEARCH ORGANIZATION ALEXANDRIA VA (1990).
13. Min, W., Spain, R., Saville, J.D., Mott, B., Brawner, K., Johnston, J., Lester, J.: Multidimensional team communication modeling for adaptive team training: A hybrid deep learning and graphical modeling framework. In: International Conference on Artificial Intelligence in Education. pp. 293–305. Springer (2021).

Promoting Explainable Feedback in Simulation-Based Training through Contrasting Case Exemplars

Caleb Vatral, Naveeduddin Mohammed, and Gautam Biswas

Institute for Software Integrated Systems, Vanderbilt University, Nashville, TN, USA

Abstract. In simulation-based training, effective in-time feedback is one of the most important factors that support high learning outcomes and trainee success. This feedback is often provided during after-action review (AAR) debriefing sessions after a training simulation and typically involves comparison of trainee actions and performance to expert baselines. With the increased use of computer-mediated simulations and low-cost sensing devices, it has become more popular to combine traditional instructor feedback during AAR with automated performance evaluations computed from collected trainee data. However, end user (both instructor and trainee) trust and understanding of these automated methods represent significant barriers to adoption and widespread use. Motivated by these barriers, in this work we explore the relationships and similarities between existing AAR structures and the tools and technologies used in explainable Artificial Intelligence (XAI). Based on this overlap, we presented a theoretical framework to integrate contrasting-case exemplar techniques from XAI into existing automated performance assessment tools to improve the transparency, trust, and usability of these automated assessments in support of AAR and continued training. We support our theoretical framework through the presentation and discussion of two case-study examples from highly distinct domains: military fire team training and nurse training. With continued research and development, the XAI integration that we propose here could help increase stakeholder trust in and acceptance of automated evaluations systems and improve the AAR process overall.

Keywords: Explainable AI, AI in Education, Simulation-Based Training, Human Performance, Automated Assessment, Human-Centered Computing, Human-Computer Interaction

Introduction

Simulation-based training environments offer many affordances that streamline the learning process for complex workplace skills compared to traditional learning environments: simulations are easily repeatable and controllable; simulations offer physical and psychological safety for otherwise dangerous tasks; data can be easily collected from simulations to provide evidence-based assessment of learners. Thus, it is not surprising that simulation-based training (SBT) has been widely adopted for training complex cognitive, psychomotor, and teamwork skills for the workplace across a wide variety of domains [1]. One of the critical factors to promoting effective learning gains in SBT is presentation of evaluation and feedback to learners after the simulation session [1, 2]. Typically, this feedback is provided during a discussion-based debriefing session, also known as after-action review (AAR), following the training. During AAR, instructors and trainees review the trainees' actions and performance to highlight and discuss both tasks that went well and areas for improvement [3, 4]. These discussions often focus around comparing trainee actions and performance to the ideal intended actions and performance as defined by field experts, and in the case of group or team training, comparing actions and performance between trainees [5, 6, 7]. However, this feedback is traditionally based on instructor observations, which are prone to errors due to the high cognitive load of observing and remembering the details of the training simulation. Because of the increasing prevalence of computer-mediated simulations, it has become increasingly available and popular to collect multimodal data about trainee actions and performance from the simulation environment and use this data to support AAR and reduce the issues of traditional instructor observations. By analyzing collected trainee data using the methods developed in the fields of intelligent tutoring and artificial intelligence (AI), it is possible to streamline the AAR process by providing automated evaluations and data evidence to supplement traditional instructor feedback and expand the discussions among trainees.

However, since these data-driven evaluations are produced using AI techniques, such as deep learning algorithms, they often suffer from the issues of interpretability and explainability that are pervasive throughout many AI models and methods [8]. While some progress has been made within the field of explainable AI, these techniques have only very recently started discussion in the education and training sector, and these issues of explainability still represent a major barrier to the widespread adoption of these data-driven AAR techniques. If stakeholders do not understand how these automated evaluations are generated, they are likely to mistrust them and advocate against their adoption. In addition, if instructors and trainees do not understand how to use the data-driven AAR technology, they are likely to revert back to traditional AAR techniques and underutilize the new technology. Motivated by these barriers, in this paper we explore how the traditional structure of AAR can be combined with techniques from the field of explainable AI (XAI) to aid in the adoption of automated evaluation techniques. Specifically, by adopting techniques from local comparison-based XAI methods, our approach extends the traditional performance metrics provided by automated AI systems to also present contrasting case exemplars that complement the existing structure and discussion of AAR. We present examples from two case studies, military fire team training and nurse training, and show the presentation of contrastive exemplars within each domain can aid in the explainability of the automated evaluations, leading to increased trust in the system and expanded discussion during the AAR.

Background

The focus of this work is on how the traditional structure and techniques used in AAR can be combined with techniques from explainable AI to lower the barriers to adoption of automated analysis techniques for SBT into AAR. To help frame this discussion, in this section we briefly review each of these three subfields and present the important background and terminology for this combination framework.

2.1 Debriefing and After-Action Review

Within SBT, one of the critical factors that mediates learning outcomes is adequate evaluation and feedback [1, 2]. Most commonly, to avoid interrupting the normal flow of events in the scenario, this feedback is provided in a discussion with instructors and peers shortly after the training session. While this discussion and feedback time goes by several different names depending on the specific domain, in the two domains that we study here, nursing education and army training, they are referred to as debriefing and AAR respectively [3, 4]. For the remainder of the paper, we will use AAR to refer to this post-simulation discussion and feedback session. During AAR, the trainees and the instructors review the events of training in order to interpret and understand the sequence of events that unfolded, how the trainees responded to the unfolding events, the links between trainee responses and expected performance, and how to improve overall performance in the future. Overall, the idea of AAR is that by analyzing trainee performance and the associated events that led to it, trainees will be given new insights for self-reflection, comparisons to peers and experts, and a better understanding of the components that make up effective performance. As a theoretical model of AAR, we adopt the model proposed by Hanoun et al. [4] which breaks down the AAR into three phases: (1) collection, (2) diagnosis, and (3) feedback. These phases are illustrated in Figure 1.

The collection phase refers to the methods and process of gathering trainee data from the simulation environment. This can take a variety of forms depending on the affordances of the simulation environment. In its most basic form, this involves an instructor or other evaluator watching the training occur and building a mental model of the events and trainee performance. Because of the very high cognitive load associated with tracking trainee performance live in this manner, training environments are often supplemented with one or more cameras that capture video of trainee actions in the simulation. In this way, the AAR can then be based on instructor observations, supported by playback and review of the video. This helps to ensure that important events are not missed by the instructor, and trainees get to review exactly what they did in the particular situation. Beyond the video capture, many modern SBT environments are designed using computer-mediated simulations, which allows collection of additional log data about simulation events. Finally, some SBT programs add additional data collection instruments that can collect a variety of other trainee data. For example, eye tracking glasses enable collection of trainee gaze and attention data, wearable biometric sensors enable collection of data about trainee stress levels, and proximity sensors enable tracking of trainee movements throughout the environment. All these additional sensors can further extend automated evaluations to generate more insights and become more robust. We review these approaches in the next section. However, at a baseline, most SBT environments include at least instructor observation and video camera.

The diagnosis phase refers to the methods and process of evaluating trainee performance and behaviors based on the data from the collection phase. These evaluations can be manual, performed by the instructor or other experts, or they can be automated, performed using AI techniques as discussed in the next section. Regardless of how the diagnosis phase is performed, it breaks down into three steps. First, *build task performance measures* refers to the process of analyzing the collected data to produce specific performance measures that are relevant in the context of the current training scenario. For example, in an emergency response domain, one potential performance measure would be the time between receiving an emergency call and arriving on the scene. Often, especially in the case of automated evaluation, performance measures are defined ahead of time; a series of domain experts will evaluate the training and develop a set of performance measures that span the requirements for effective performance. Then, immediately after training when the data becomes available, the predefined performance measures are computed from the collected data. Second, *assess task performance* refers to the process of evaluating the performance measures against a set of pre-set constraints. Returning to the emergency response example, our response time performance measure could be compared against a pre-defined temporal window that defines acceptable performance. Response time within this window would be assessed as satisfactory performance, while outside this window would be assessed as unsatisfactory or deficient performance. Finally, *compare actual performance to intended* refers to the process of comparing the trainee actions and performance measures to expert or expected performance. This comparison to intended performance provides trainees with insights and actionable changes that they can make to their task execution to improve performance in the future during the feedback phase of AAR.

Finally, the feedback phase refers to methods and process of presenting the results of the diagnosis phase to the trainees, so that they can improve their performance in the future. This phase breaks down into two primary steps. First, *review comparison results* presents the results of the comparison step of the diagnosis phase to the trainees. As the trainees review the differences between their own actions and performance compared to optimal performance, they can begin to discuss the causal relationships that lead to the actions, decisions, and behaviors that the trainees exhibited. Fundamentally, this step is about answering the question, “why was my performance different than the intended performance?”. After this, *discuss performance* further opens the discussion to focus on what the trainees can do to improve performance in the future. After reasoning about why they performed differently than intended, trainees can begin to generalize these cause-and-effect relationships to determine what they can do next time to perform better.

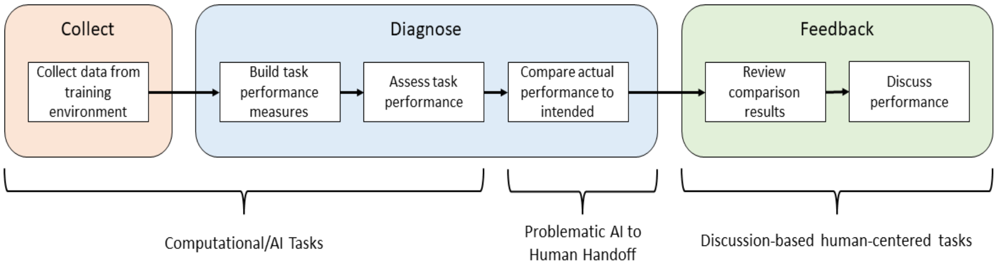
Collect
Collect data from training environment
Diagnose
Build task performance measures
Assess task performance
Compare actual performance to intended
Feedback
Review comparison results
Discuss performance
Computational/AI Tasks
Problematic AI to Human Handoff
Discussion-based human-centered tasks

Fig 1. The three-phase AAR framework adapted from Hanoun et al. [4] (section 2.1) and labels of tasks performed using either automation/AI or human-centered tasks (section 2.2).

2.2 Automated Performance Evaluation

With the advent of modern computing technology, many SBT environments have begun to adopt computer-mediated simulations as a major component of their architecture [3]. These computer-mediated simulations offer higher fidelity to trainees, as they can adapt realistically to trainee actions in the environment, but they also offer additional benefits: (1) the ability to create variations of a scenario in a simple cost-effective manner to enrich the training process; and (2) the ability to streamline data collection for analysis of trainee performance and behaviors to streamline AARs. Computer simulations are capable of keeping detailed logs of trainee actions in the environment, thus reducing the cognitive load on instructors and decreasing the potential for missed data. In addition to these logs, computer simulations are often capable of synchronizing the capture of video data and other sensors with the simulation logs, providing a more complete picture of the simulation to review during AAR. With this increase in the volume and accuracy of simulation data, along with advances in the fields of AI and intelligent tutoring, it has become increasingly accessible to automate performance evaluation of trainees for use during AAR [9, 10, 11].

While a variety of analytic techniques and models can be used to perform automated performance evaluation, the design and computation of performance assessment largely follows closely to the diagnosis phase of AAR described in the previous section. First, before the simulation begins, a group of domain experts design a set of performance metrics that measure and span the concepts targeted during the training [9]. When performing automated assessments, it is important to also consider how the performance metrics will be calculated when defining them. Designers must ensure that each performance metric has an associated algorithm capable of calculating the metric from the collected data. These computation algorithms often utilize tools and methods from signal processing, machine learning, and AI, and they are created through collaboration between simulation domain experts, the instructors, and experts in computational fields, such as computer scientists [9, 10]. Next, after a simulation completes, the data collected from that simulation is passed into the previously designed algorithms to compute the performance measures. However, this is where most automated performance evaluation systems stop. The computed performance metrics are given back to instructors and trainees, often in the form of analytic dashboards, and it then becomes the responsibility of these stakeholders to continue the AAR process into the comparison and discussion phases.

However, this passed responsibility can often cause breakdowns in the AAR process. First, if instructors and trainees do not understand the computational decision making – i.e., how the performance metrics were computed – they likely will not put much faith in the results. This is especially the case when the automated assessments contradict the instructor's viewpoints and feedback in some way. This lack of transparency from the algorithms leads to a lack of trust by end users, and ultimately, this will cause end users to abandon the system. Second, if instructors and trainees do not understand how to use the results of the automated evaluations, then they will likely significantly underutilize the system. Often, instructors are provided with complex analytics presented in dashboards that the instructors were not trained on, and that were created without first performing a needs analysis. This can lead to the instructors using the automated evaluations at a cursory level but missing the greater detail and new insights that these tools can provide. To remedy these issues, we propose integrating some of the techniques from XAI in order to extend the automated performance assessment systems into the next stage of the AAR process to provide comparisons and promote discussion.

2.3 Explainable AI

As systems built using the tools and technology from the field of AI become more pervasive, it becomes increasingly important to ensure end users have a sense of trust and understanding of these systems. Within the education and training sector, this is no different; learners and instructors must be able to understand how AI systems work, how the systems affect them, and whether the systems are trustworthy [12]. This need for understanding and trust has led to increased research and development in *XAI*, which refers to the tools and techniques used to introduce trust between end users and AI systems by providing methods of understanding how the AI works and how it makes its decisions [8].

Within XAI, there are several unique goals that each focus on a different aspect of the overall explainability, and research in XAI typically focuses on improving a small subset of these goals. In this work, we adopt the XAI model developed by Fiok et al. [8], which breaks down XAI into seven goals. These goals are summarized in Table 1. From this set of goals, in this work, we focus specially on *transparency* and *trust. Transparency* in XAI refers to how well end users understand the model's decisions and decision-making process. Within the context of SBT and automated performance evaluation, these decisions represent the performance metrics associated with each trainee, and the decision-making processes represent the algorithms used to calculate these performance metrics. As previously discussed, if instructors and trainees do not understand how the performance metrics were calculated and what factors contributed to them, then they will likely disregard the automated assessments and rely only on instructor observations. *Trust* in XAI refers to how confident users are in working with the AI system and using its outputs. Within the context of SBT and automated performance evaluation, this means that instructors and trainees are able to use the evaluations generated by the system to promote meaningful discussions during AAR and to take the insights that they gain from the AI evaluations to change their behaviors in future training to improve performance. If users do not understand how to use the automated evaluations that the system generated and do not understand how these evaluations apply to their own training, then they will only use the system at a very surface level and not gain the full benefits that it offers.

Table 1. The seven goals of explainable AI, summarized from Fiok et al. [8]

XAI Concept	Description
Transparency	Do end users understand the model's decisions?
Reliability	Is the system robust against changes to parameters and inputs?
Usability	Is the system capable of providing a safe and effective environment for end users?
Trust	How confident are the users in working with the system and using its outputs?
Fairness	Are the model's decisions free from bias and fair over protected groups?
Privacy	How does the system protect sensitive user information?
Causality	Do small changes to the inputs of the system produce the expected changes to the output?

Several classes of methods have been developed to address explainability in AI systems. Khosravi et al. [12] categorized these methods into two major classes: global approaches and local approaches. Global approaches to XAI focus on explaining the overall model prediction, typically by examining how the input features affect predictions. One example of this global approach is to examine feature relevance. By comparing perturbations in the input features to an AI model, feature relevance systems can assign a score to each input feature based on how much that feature contributes to the overall decision. These global models have unique advantages, especially for system designers, as it allows them to examine the overall model behavior to ensure reliability, causality, and fairness. On the other hand, local approaches to XAI focus on explaining specific prediction instances rather than the overall model. These local approaches have unique advantages as well, especially for end users, who are generally less concerned about overall system behavior and more concerned about how the system affects their individual use cases. These local approaches are generally based on examining specific exemplars of the model, and can help end users reason about transparency, usability, and trust.

Within local approaches to XAI, one technique which is particularly applicable to SBT and team training is comparison-based approaches. In comparison XAI, specific instances are used to locally explain the outcomes of other instances by comparing and contrasting the input features and final decision outputs for each of the instances. For example, many automated plagiarism detection systems present comparisons between two texts as evidence to explain why they predicted plagiarism [9]. Noting the similarity of these comparison-based methods to the diagnostic and feedback phases of AAR, in this work we focus on the application of comparison-based XAI approaches to automated performance evaluation systems. In the next section, we will discuss these similarities in more depth and present the theoretical framework for our approach.

Theoretical Framework

In this section, we describe the proposed framework to integrate techniques from XAI with the three-phase model of AAR. Specifically, we utilize the similarities between comparison based local XAI approaches and the structure of the diagnostic phase of AAR to combine the two techniques into a singular framework that extends the existing structure of automated performance evaluation systems for SBT. This new theoretical framework is illustrated in Figure 2.

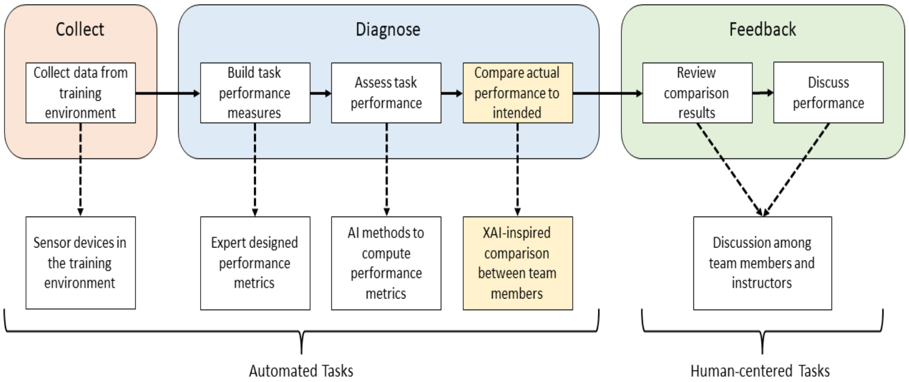
Collect
Diagnose
Feedback
Collect data from training environment
Build task performance measures
Assess task performance
Compare actual performance to intended
Review comparison results
Discuss performance
Sensor devices in the training environment
Expert designed performance metrics
AI methods to compute performance metrics
XAI-inspired comparison between team members
Discussion among team members and instructors
Automated Tasks
Human-centered Tasks

Figure 2. The theoretical framework to combine existing 3-phase AAR structures with comparison-based XAI techniques to improve the transition between automated and human-centered tasks.

The new framework follows largely the same structure as existing techniques for automated performance evaluation. First, in the collect phase, data is collected about trainee actions in the simulation environment using a variety of sensor devices (cameras, log files, etc.). Then, this data is passed to the diagnostic phase, where it is used to compute a set of expert-designed performance metrics. As previously described these metrics are designed to evaluate and span the complete range of tasks that trainees are expected to perform in the given context. The metrics are computed via AI algorithms which process the raw trainee data collected from the environment. However, at this point, the proposed framework deviates from the existing methods. In current systems, these performance metrics are typically provided back to instructors and trainees at this point, and it is the responsibility of these end users to interpret the automated evaluations and use them to help inform comparison to intended performance and further discussion during the feedback phase. In the new framework, this last step of the diagnosis phase (highlighted in yellow in Figure 2) is instead performed by the AI system using comparison based local XAI techniques. Instead of transitioning from automated tasks to human-centered tasks during the diagnostic phase, our new framework allows the transition to occur between the diagnostic and feedback phases, thus completing the automation of collection and diagnosis and improving the transparency and actionability of the performance metrics.

This critical last step in the diagnosis phase of AAR represents the comparison of actual trainee performance to the theoretical intended performance (exemplars as provided by experts). This comparison is nearly exactly the same technique used in XAI to explain local decision making. By comparing the AI model's outputs, in this case the trainee performance metrics, in one instance to another instance (e.g., examples of better trainee performance, or examples of expected performance as defined by the experts and instructors), end users will be able to better understand and evaluate how the performance metrics were computed, which, in turn will help them gain better insights on how to improve performance in the future. To perform this XAI-inspired comparison, we rely on contrasting an individual trainee's data with two other exemplars that we briefly outlined above: expert performance and peer performance.

Comparison to expert performance is a cornerstone of traditional feedback in AAR. Typically, expert instructors watch the trainees' actions in the simulation, and then highlight things that the instructor would have done differently if they were performing the exercises. To mimic this idea computationally in the AI system, we setup an expert-model exemplar when initially designing the performance metrics and the training scenarios. For example, returning to the example of the emergency response domain and our response time metric, the comparison methods might examine the path taken by ambulance from dispatch to the emergency scene. When initially designing the training scenario, the instructor (or other expert designer) would specify the optimal or expert path for the ambulance to take that would optimize the response time. Then after the trainees complete the simulation, this expert path would be compared to the path taken by the trainees. If the trainees took an unnecessarily long path to the emergency scene, or ignored information on congestion along their chosen path, then their response time will generate scores that would fall in the inadequate interval. However, in addition to the raw score, trainees could be shown the differences between their path and the expert path, with explanations from the expert as to why they chose a particular path. The overall idea is that trainees not only get a score, but also an explanation why that score was low compared to an expert, thus promoting further discussion during the feedback phase on how they may improve their decision making to choose better paths in the future.

Beyond comparison to expert exemplars, we can also perform comparison between individual trainee performances. By directly comparing trainee performance, we can help to highlight differences in the decision-making processes between people at the same level of expertise. Computationally, we can choose trainee exemplars to compare using a further than K-nearest approach. First, with a further than K-nearest neighbors approach, the AI system would compare the performance metric scores of each pair of trainees, and select a set of contrasting exemplars (i.e. one score high and one score low). In this way, the high-scoring trainees provide examples of how trainees can incrementally improve performance, and not just have to jump to the expert-defined “best” performance. Further, the differences between the high-scoring students and low-scoring students can help explain using the scoring scheme to determine how low-scoring trainees might change their actions to improve performance in the future. Second, with a K-nearest neighbors approach, the AI system would again compare the performance metric scores of each pair of trainees, but this time select similar examples (i.e. both high scoring). This type of comparison may offer a different type of insight into trainee behaviors. For example, if two trainees both score well but had highly disparate actions to achieve their goals, then contrasting these exemplars could help illustrate the different strategies to approach the same problem while still maintaining task success. In the next section, we will show examples of both trainee-to-expert and trainee-to-trainee comparison on two case study domains, and how presenting these contrastive exemplars can help promote discussion during the feedback phase of AAR.

Case Study Examples

To demonstrate the approach presented in this paper, we utilize two case studies from different domains and show that the same methods apply across both domains to establish the generality of our approach. This section we will briefly review the two studied domains, military fire team training and nurse training, and discuss the unique aspects and performance metrics for each domain.

4.1 Military Fire Team Training

The fire team training took place at the Ft. Campbell US Army installation and focused on the Enter and Clear a Room (ECR) dismounted battle drill. Training of the drill was conducted using the Squad Advanced Marksmanship Trainer (SAM-T) system, which is a mixed-reality environment for simulating live-fire weapons training drills. For the ECR drill, the SAM-T is setup using three screens configured in a U-shaped arena. The fire team moves around in the space created by the arena to simulate moving through a physical room, and they interact with a digital simulation projected onto the three screens. The digital simulation is interfaced with the soldiers’ weapons, allowing them to realistically fire at on-screen entities. Other interactions with the digital simulation are mediated by on-site instructors, who watch the team’s actions and update simulation parameters in real-time. In ECR, the team is tasked with entering an unknown room, neutralizing any enemy combatants, and securing any civilians. To perform our automated evaluation techniques, we collected data from the simulations including two cameras, which captured audio and video of the soldier actions in the physical space, simulation log files, which captured digital simulation events such as weapons firing and digital entity movements. For more information about the fire team training, see [9].

For the case study in this paper, we focus specifically on one of the performance metrics used to evaluate this domain: *move-along-walls*. The *move-along-walls* metrics evaluated whether the soldiers in the fire team stayed close to the walls of the room while entering. From a doctrine perspective, soldiers should move along the walls of the room in order to minimize potential blind spots at their backs. To compute this metric, we utilize computer vision-based motion tracking techniques applied to the video data to track the soldiers’ movements throughout the room. In each frame of video, we can measure the distance between the soldier and the closest wall. Then, we average this distance across the entire simulation and normalize it between zero and one based on

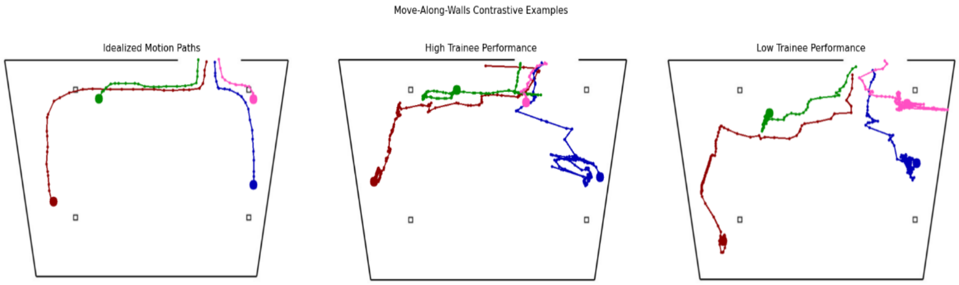

Figure 3. Motion paths illustrating the move along walls metric for 3 contrasting examples

a maximum threshold distance to obtain the final performance metric score. To perform exemplar comparison, we compare the final scores of the soldiers and present the movement paths that they took through the room against both expert examples and peer performance.

Figure 3 shows the motion paths of the fire team throughout the simulation room for three contrasting examples of the *move-along-walls* metric. First, the left-hand image in Figure 3 shows an example of an expert motion path for the team, representing an idealized performance based on Army training doctrine. The middle image in Figure 3 shows an example taken from the case study where the fire team performed high on the *move-along-walls* metric. Finally, the right-hand image in Figure 3 shows an example taken from the case study where the fire team performed low on the *move-along-walls* metric. In this case, we see two separate instances of the contrastive exemplar technique. First, trainees can compare their actual performance to idealized performance to identify issues with their current performance and discuss ways to improve in future training. For example, in the high trainee performance instance, the team might notice and discuss that the soldier represented in blue cut directly to their destination rather than moving right to the wall and then down as shown in the expert example. Given this issue apparent in the visualization, the team understands why they received a low performance score (i.e., AI transparency), as well as discuss how they can rectify this issue in future training (i.e., AI trust). Second, trainees can compare their performance to their peers. In this case, the team acts as their own peer exemplars. By viewing the comparison between the high-performing instance and low-performing instance, the team can begin to recognize what was different between the two instances that may have contributed to the gap in performance evaluations. In this example, the team might notice that in the high-performing instance, the soldiers represented in green and red both moved closely along the back wall, while in the low-performing instance, these soldiers penetrated deeper into the room on their initial entry before moving across. Both examples, expert comparison and self (peer) comparison, provide unique insights into the team's overall performance and can help them to understand why they scored how they did and increase discussion about how to improve in the future.

4.2 Nurse Training

The nurse training program took place in a simulated hospital room equipped with standard medical equipment which has been modified to interface with a high-fidelity patient manikin. In each simulation, nurses enter the hospital room with the goal of performing a routine clinical assessment of the patient. However, during the assessment, the patient exhibits symptoms of a new medical condition. After observing the new symptoms, the nurse is responsible for evaluating the patient to determine a potential diagnosis, performing any relevant procedures to stabilize the patient's condition, and contacting the medical provider to update them on the new condition and receive new orders. All the participants for this study were undergraduate (BSN) level nursing students in their first year of coursework, and the simulations we study in this paper were part of the students' normal degree requirements. No changes to the content of the simulations were made by the researchers. To perform our automated evaluation techniques, we collected data from the simulations including two overhead video cameras which captured video and audio of the entire hospital room, eye tracking glasses which captured egocentric video and audio of the participants along with their gaze coordinates, and log files which captured the manikin's simulation parameters (vital signs, etc.). For more information about the nurse training simulation, see [10].

For the case study in this paper, we will focus specifically on one of the performance metrics used to evaluate this domain: *dialogue-space-distribution*. The *dialogue-space-distribution* metric evaluates how well the nurse keeps in communication with the patient throughout the simulation. From a best-practice perspective, nurses should be in persistent communication with their patient throughout the course of diagnosis and treatment. The nurses should be in communication with the patient about both what the patient is feeling to monitor any changes, and about what the nurse is currently doing to ensure that the patient is well informed and feels safe. This metric examines the percentage of dialogue that the nurse performs while on each side of the patient's bed. Since the nurses must move between each side of the bed to perform various clinical tasks (e.g., reading the chart vs taking vital signs), this distribution of dialogue can help to highlight specific tasks where the nurse might get distracted and forget to keep in communication with the patient. To perform exemplar comparison, we compare the distributions of dialogue between different nurses (peer comparison).

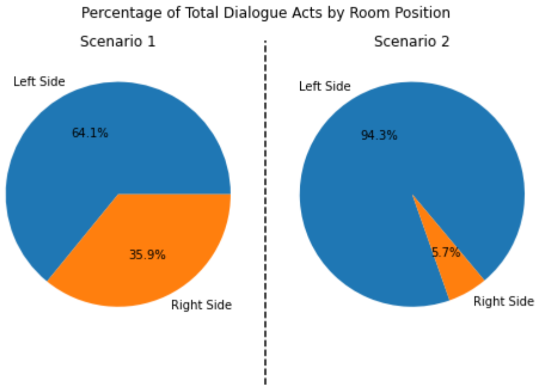

Figure 4. The *dialogue-space* distribution for two contrasting nurse trainees

Figure 4 shows the *dialogue-space* distribution for two of the nurses from the case study which represent peer contrastive exemplars. Nurse 1 (Figure 4, left) performed 64.1% of her dialogue on the left side of bed and 35.9% of her dialogue on the right side of the bed. Contrastively, nurse 2 (Figure 4, right) performed 94.3% of her dialogue on the left side of the bed and only 5.7% on the right side. In this case, while both nurses performed more of their dialogue on the left side of the bed, there is a clear difference in the overall dialogue behaviors of the two nurses, making these contrasting exemplars ideal candidates for the AI system to highlight for further discussion during the feedback phase of AAR. During this feedback phase, the instructor could utilize the difference between these distributions to engage a discussion about best practice for dialogue with the patient. Was there a specific reason that nurse 2 did not engage in much dialogue when on the right side of the bed? What are some of the common causes that lead to lack of communication with the patient? How can nurses ensure that they communicate with the patient sufficiently often so that patients do not feel they are being ignored? What strategies were used by nurse 1 to keep these best practices in mind that might be useful to other trainees in the group? By presenting the trainees with these contrasting peer examples, they can use the AI-generated evaluations to learn from one another and contribute to a larger discussion about how to improve everyone's skills.

Conclusions and Future Work

In this paper, we explored the overlap between existing AAR structures and the tools and technologies used in XAI. Based on these similarities, we presented a theoretical framework based on contrasting cases, to integrate XAI techniques into existing automated performance assessment tools to improve the transparency, trust, and usability of these automated assessments to support AAR and continued training. Our new approach leverages local example-based XAI techniques to present end users with the automated performance assessments in an interpretable manner. By presenting instructors and trainees with contrasting examples of performance in the training domain, we can improve the users understanding of how the performance metrics were computed and promote constructive discussions for how trainees might improve performance in the future. Contrastive exemplars can provide comparison between trainee performance and an idealized expert performance, but are also particularly applicable for team training domains, where comparisons can be made between peers at the same level of expertise to promote learning from each other. We showed some initial examples of expert and peer contrastive performance to generate explainable feedback in two distinct domains – military fire teams and nurses – and showed how presenting these contrastive exemplars can promote AI transparency and improved discussion during AAR.

While the initial results from the two presented case study examples are promising, this work represents ongoing research in its early stages and requires future work to further develop and validate this approach. In particular, future work will need to focus on automated methods for selecting useful contrastive contextualized examples to generate meaningful explanations. While we have adopted a preliminary exploration approach using the K nearest neighbor algorithm (within the K-nearest bounds and outside of the K-nearest bounds, discussed in Section 3), the case study examples presented here were hand selected to illustrate our overall goals of the framework. We will continue to develop these automated contrastive selection techniques, and extend them, for example by overlaying multiple similar examples to create a generic case that illustrates good and bad performance of a particular concept or technique, so that these examples can support end users without a lot of human intervention in the future. In addition, significant future work will be devoted to usability testing of this approach. Through interviews with stakeholders and instructors, as well as studies that incorporate this new feedback mechanism during live training events, we will fully refine and validate both the framework and the tools to present the results. With continued research and development, we hope that contrastive exemplar techniques such as these can increase stakeholder trust in automated evaluations systems and help to improve the AAR process overall.

Acknowledgements

This work represents independent research supported in part by US Army DEVCOM Soldier Center Award W912CG2220001, as well as equipment and funding from the Vanderbilt LIVE initiative. The views expressed in this paper do not necessarily reflect the position or policy of the United States Government, and no official endorsement should be inferred.

References

P. Ravert. An Integrative Review of Computer-based Simulation in the Education Process. CIN: Computers, Informatics, Nursing, 20.5 (2002): 203-208. doi: 10.1097/00024665-200209000-00013

A. Gegenfurtner, C. Quesada-Pallarès, and M. Knogler. Digital simulation-based training: A meta-analysis. British Journal of Education Technology, 45 (2014): 1097-1114. doi: 10.1111/bjet.12188

T. L. Sawyer and S. Deering. Adaptation of the US Army's After-Action Review for Simulation Debriefing in Healthcare. Simulation in Healthcare: The Journal of the Society for Simulation in Healthcare, 8.6 (2013): 388-397. doi: 10.1097/SIH.0b013e31829ac85c

S. Hanoun and S. Nahavandi. Current and future methodologies of after action review in simulation-based training, in: Proceedings of the 12th Annual IEEE International Systems Conference, Vancouver Canada, 2018.

P. G. Sidney and M. W. Alibali. How do contrasting cases and self-explanation promote learning? Evidence from fraction division. Learning and Instruction, 40 (2015): 29-38.

D. L. Schwartz, C. C. Chase, M. A. Oppezzo, and D. B. Chin. Practicing versus inventing with contrasting cases: The effects of telling first on learning and transfer. Journal of educational psychology, 103.4 (2011): 759.

D. L. Schwartz and J. D. Bransford. A time for telling. Cognition and instruction, 16.4 (1998): 475-522.

K. Fiok, F.V. Farahanu, W. Karwowski, and T. Ahram. Explainable artificial intelligence for education and training. Journal of Defense Modeling and Simulation: Applications, Methodology, Technology, 19.2 (2022): 133-144. doi: 10.1177/15485129211028651

C. Vatral, G. Biswas, and B. S. Goldberg. Multimodal Learning Analytics Using Hierarchical Models for Analyzing Team Performance, in: Proceedings of the 15th International Conference on Computer Supported Collaborative Learning. International Society of the Learning Sciences, Hiroshima Japan, 2022, pp. 403-406.

C. Vatral, G. Biswas, C. Cohn, E. Davalos, and N. Mohammed. Using the DiCoT framework for Integrated Multimodal Analysis in Mixed-Reality Training Environments. Frontiers in Artificial Intelligence (2022): In Press. doi: 10.3389/frai.2022.941825.

G. Biswas, R. Rajendran, N. Mohammed, B. S. Goldberg, R. A. Sottilare, K. Brawner, and M. Hoffman. Multilevel learner modeling in training environments for complex decision making. IEEE Transactions on Learning Technologies, 13.1 (2020): 172-185. doi: 10.1109/TLT.2019.2923352

H. Khosravi, S. Shum, G. Chen, C. Conati, Y.S. Tsai, J. Kay, S. Knight, R. Martinez-Maldonado, S. Sadiq, and D. Gašević. Explainable artificial intelligence in education. Computers and Education: Artificial Intelligence, 3 (2022).

Machine Learning Approaches for Assessing Team Learning, Performance and Proficiency

Robert A. Sottilare, Ph.D.[0000-0002-5278-2441]

Soar Technology, Inc., Orlando, FL 32817, USA
bob.sottilare@soartech.com

1 Introduction

There are major challenges in modeling how teams learn, determining methods to validate the level of their learning (knowledge and skills), and predicting their potential for success which indicates their transfer of skills to new training contexts and operational (work) environments. In this paper, our goal is to identify a process that is congruent with the training and learning science literature that could lead to a practical data management process where machine learning (ML) is used to accurately classify current and predict future team states, specifically team learning, team performance, and team proficiency. An accurate, consistent, repeatable data management process would have great value in modeling the impact of training (both traditional and adaptive training), but especially adaptive training where team assessments could be used to tailor events within future training scenarios, identify remediation and feedback for use in after action reviews, and generate recommendations for future training consistent with organizational development goals.

Adaptive training is delivered and monitored by a technology called an adaptive instructional system (AIS) that accommodates (tailors) training and educational experiences based on individual differences to facilitate learner knowledge acquisition [1,2] and guide tutoring activities that exercise skills defined by a set of learning or training objectives [3]. AISs use artificial intelligence (AI) and other advanced technologies to help people learn more effectively and efficiently [4]. Historically, it has been much more difficult to model teams and determine predictors of team learning, performance, and proficiency. This requires a rich dataset and a technical approach that is both flexible and scalable to support teams within both large and small organizations.

Towards our goal of supporting adaptive team training, we provide a concise set of definitions for team learning, team performance, and team proficiency which will be expanded upon in Section 2 of this paper. *Team learning* is the collective acquisition of knowledge and skill by a group of individuals with different roles and responsibilities within the team. *Team performance* is the measurement of a group's ability to successfully complete specified tasks, actions, or functions under a set of conditions in a controlled environment where measurement is possible. *Team proficiency* is an assessment of a group's mastery of a set of tasks that represent the required operational (working) knowledge and skills required to function in a job or position.

This paper also describes a data management process for assessing the value of measures of team learning, performance, and proficiency as part of an ML approach. ML is a subset of AI that uses data to classify/categorize, make inferences about relationships or trends or make predictions. Specifically, we will be discussing the use of an ensemble model approach that includes various types of models including deep neural networks (DNNs). DNNs have several advantages over classical ML methods (e.g., linear regression, clustering, decision trees, or nearest neighbor techniques) including the ability to automatically generate features, the ability to perform well with unstructured data, improved self-improving model capabilities, and most importantly scalability [5]. The use of DNNs within a stacked (hierarchical) ensemble model with other ML techniques will allow AIS developers to reduce risks associated with weak individual models and overfitting where models are so tightly coupled to the dataset used to create the model that they are unable to generalize well to new data.

The ability to detect relationships between features and critical outcomes using DNNs can identify significant measures of assessment for all three of our outcomes of interest: learning, performance, and proficiency. The result is that our ML models will be able to detect strong associations that are identified using a relatively small set of critical features. Details of the data management process including model building and testing will be described in Section 4 of this paper. In the next two sections, we compare and contrast team learning, performance, and proficiency (Section 2) and extract candidate features from the team assessment literature (Section 3).

2 Comparing Learning, Performance and Proficiency

In this section, we define, compare, and contrast learning, performance, and proficiency to lay the groundwork for determining critical measures for team learning, performance, and proficiency processes. These measures will be used later in this paper to identify ML features to infer current learner states and predict future learning, performance, and proficiency trends.

2.1 Individual and Team Learning

Learning is any increase in knowledge or skills that results in a relatively permanent change in knowledge or behavior as measured through assessments [6]. *Assessments* are opportunities to determine which specific knowledge or skills were acquired [7], but may also be opportunities to identify deficiencies that may need to be addressed during future learning experiences (e.g., training or education). In addition to measuring individual or team learning, assessments are also used to influence decisions about the curriculum, instructional content, and content delivery methods with the goal of improving learning programs. Assessments may be categorized as:

- *diagnostic* - used to determine current knowledge or skill prior to a learning activity; an example of a diagnostic assessment is a pre-test
- *formative* - formal and informal assessments during learning activities provide evidence for modifying teaching methods or learning activities; examples of formative assessments are checks-on-learning and concept mapping activities
- *interim* - tests administered at intervals between formative and summative assessment to check learner comprehension and guide future instruction; examples of interim assessments are essays or projects
- *summative* - assessment activities where the focus is to determine whether the goals (outcomes) of the learning program were achieved; examples of summative assessments are final exams or capstone projects

Both individual and team learning are assessed using measures to determine progress toward learning or training objectives. Measures are task or domain dependent in that measures of assessment for cognitive tasks may differ from those used for psychomotor or team tasks. Team learning differs from individual learning in that it requires individual team members to interact and coordinate to successfully achieve defined objectives. Whereas collaborative learning is “a situation in which two or more people learn or attempt to learn something together” [8], teamwork is the “coordination, cooperation, and communication among individuals to achieve a shared goal” [9] and involves a domain-independent assessment of the team’s interactions.

2.2 Individual and Team Performance

Performance is the application of domain knowledge and skill to efficiently accomplish one or more actions, tasks, or functions (e.g., problem solving, optimizing decisions) [10]. The relationship between knowledge and skill is captured in the definition of *skill*, which is both learned and applied abilities that use one's knowledge effectively in the execution or performance of a task [11]. Performance can be measured or observed during instructional or work (operational) experiences. Individual and team performance usually fluctuates based on workload, skill decay, or changing conditions in the instructional or work environment. For team-based activities, performance may fluctuate due to teamwork factors such as leadership, team cohesion, coordination, effective communications, or other factors .

2.3 Individual and Team Proficiency

Proficiency is a measure of acquired knowledge and skills along with attitudes and behaviors that lead to the ability to do a task successfully or efficiently [12]. Both proficiency and *competency* involve the application of skills to the performance of a specific task or set of tasks. However, the difference between proficiency and competency is that competency refers to a set of essential skills that are required for performance, and proficiency implies a level of mastery of these essential skills. For example, a learner may be competent as a land navigator in that they possess all of the essential skills (e.g., map reading, compass use, and landmark identification) required to move from one position to another across the landscape. While all of the required skills are present, the learner may be barely proficient in each of them, or they may be better at some than others, or they may be fluent (highly proficient) in all of them. The same may be said of team proficiency as the individuals composing a team bring different skillsets (strengths and weaknesses) relative to the goals being pursued by the team. While it is not just a simple matter of adding the proficiencies of individual team members to determine team proficiency, it is an indicator of missing skills within the team, and may be used to understand the team's probability of mission success.

Proficiency may be measured at various levels (e.g., Army competency standards may equate to "at expectation" within an AIS). Part of proficiency measurement is understanding how various training scenarios compare to operational processes and standards. Some training is incremental in that it pulls in only part of the challenges (conditions) that a soldier might experience in an operational/working context. An expert will be able to efficiently execute a set of tasks associated with their domain of expertise. So, measures of efficiency (speed, agility, accuracy) are often good features (inputs) to ML models made to classify or predict proficiency. Cumulative experience (including training) supports the notion that a number of hours of deliberate practice are needed to harden skills, and of course time is a factor in skill decay. Finally, measures are needed to understand how readily a soldier might be able to apply knowledge and skill in new contexts, scenarios of higher difficulty, and sets of conditions or constraints that impart stress.

2.4 Relationships of Learning, Performance, and Proficiency

Since learning is the acquisition of knowledge and skill, and knowledge and skill are requisite precursors for performance, then learning and performance are related. Temporally, learning occurs over a period of time and multiple instructional experiences. While, learning results in permanent changes in behavior, it is an indicator and a predictor of performance, it is not a guarantee of consistent performance which can vary even among domain experts. Past performance is just one of many indicators predicting success in future scenarios composed of events and varying sets of conditions. Performance also demonstrates learning and proficiency.

A simplified model of the relationship between learning, performance, and proficiency identifies learning as a prerequisite for performance, and performance as a prerequisite for proficiency (Figure 1). It also notes that performance is required to demonstrate both learning and proficiency, and is therefore central to assessing and predicting learning, and proficiency. While this proclamation passes the common-sense test, the actual relationships are much more complex. For example, there are temporal variables that influence assessment models for team learning, performance and proficiency. Learning occurs over a series of instructional events. Performance, which reflects the team's ability to apply acquired knowledge and skills, occurs within a single controlled event. Finally, proficiency involves achievement of successful performance across multiple operational events where the team demonstrates a level of expertise.

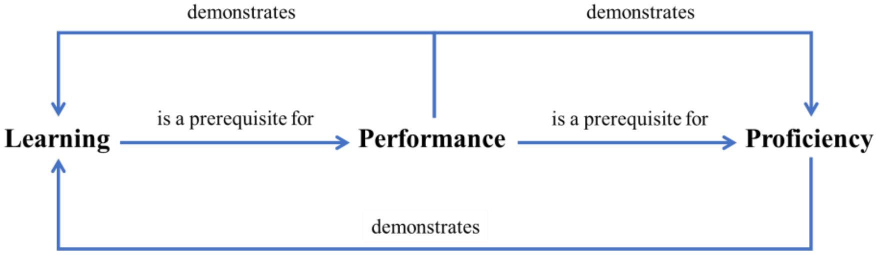

Fig. 1. Simplified model of the relationship between learning, performance and proficiency

Given our understanding of the relationship and influences of learning on performance and proficiency, and the use of performance measures to validate learning and proficiency (mastery of skills), the next step in this process of examining team learning, performance, and proficiency is to begin identifying specific antecedent variables and predictors. An antecedent variable is present before the independent and dependent variables under study and moderates or helps explain the relationship of other variables. The reason to identify these variables in the literature is that they are often feature engineering candidates in the data management process and are useful for developing ML models, and supporting our goal of accurately assessing team learning, performance, and proficiency across various domains of instruction.

Performance is a demonstration of proficiency. Proficiency is demonstrated (applied) knowledge and skill that predicts, but does not guarantee future performance (the ability to use/apply knowledge). Proficiency may also be viewed as the potential for performance or potential for successful completion of a task, and it is represented by varying levels of skill development while competency signifies that the learner has achieved at least the minimum threshold of proficiency which is usually defined by a measurable standard. Task proficiency may include categories or levels such as no proficiency, elementary proficiency, limited working proficiency, professional working proficiency, full professional proficiency, and expert proficiency which usually contain qualifiers that describe knowledge, skill, and abilities at each level. While proficiency is a predictor of future performance, it is not a guarantee since the conditions under which tasks are performed can vary and influence performance. For example, a learner who has mastered skills for a land navigation task in an urban environment may encounter new conditions (e.g., heavily forested areas) that influence their performance. New conditions might also include changes in the condition of the learner (task performer) that can influence their performance – lack of practice, lack of energy, lack of sleep, lack of confidence, skill decay or the inability to transfer skills from prior experiences to a new experience. This is why performance fluctuates.

3 Examining Measures of Team Assessment

In reviewing the literature, our goal was to identify attitudes, behaviors, and cognition variables that significantly influence team learning, performance, and proficiency outcomes. To this end, we began with learning as a precursor for successful performance, competency (essential skills) development, and proficiency (mastery level of essential skills), but we also sought to understand how domain learning is represented in behaviors and assessed (Figure 2). We developed our ontology from this point of view.

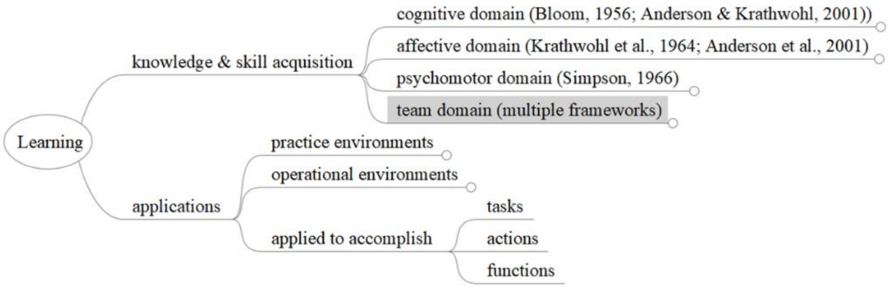
Learning
knowledge & skill acquisition
cognitive domain (Bloom, 1956; Anderson & Krathwohl, 2001))
affective domain (Krathwohl et al., 1964; Anderson et al., 2001)
psychomotor domain (Simpson, 1966)
team domain (multiple frameworks)
applications
practice environments
operational environments
applied to accomplish
tasks
actions
functions

Fig. 2. Domain Taxonomies for Learning Activities and Assessments

We began by identifying domain taxonomies (cognitive, affective, psychomotor, and team) for learning activities and assessments defined in the literature. Unlike the cognitive, affective and psychomotor domains, there was not a definitive and comprehensive single source for the team domain, but there were three candidate frameworks that helped us identify team learning activities and assessments, and relationships to team performance, and proficiency. The first was the temporally-based framework for team processes [13] addressed as a series of input-process-outcome cycles that represent attitudes, behaviors and cognitive processes during transition phases (involving planning, setting objectives, and defining strategies) and action phases (involving activities). The second framework, a taxonomy of teams, team tasks, and tutors [14], identified team performance context in the form of team attributes, team tasks, and team skills. The third framework, identified attitudes, behaviors and cognition related to cooperation, coordination, communication, cognition, coaching, conflict, and conditions to assess teamwork [15].

Based on this composite team modeling approach, we identified activities (team behaviors and processes) and measures of assessment associated with team learning, performance, and proficiency (Figure 3). The goal is to develop a list of candidate features that can be tested within models that are created and trained within a ML pipeline process used to classify/predict team performance states, and identify trends within the training population.

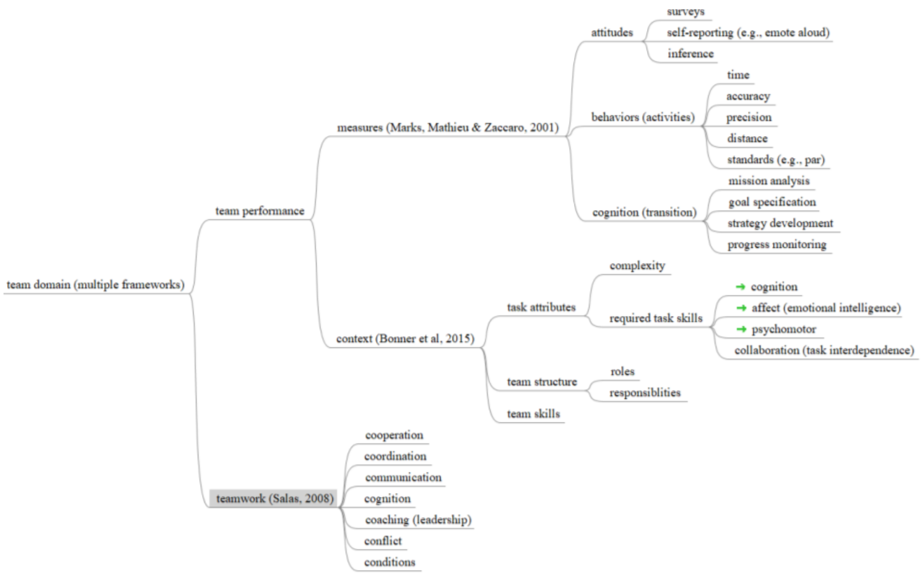

Fig. 3. A Team Learning, Performance & Proficiency Framework

Based on the literature and our initial investigation of data relationships, candidate features for classifying and predicting team learning, team performance, and team proficiency should include:

- team learning classifiers and predictors – team performance and proficiency measures including transitions (evidence of mission analysis, planning, goal specification, strategy development and adaptation, progress toward learning objectives, and markers related to teamwork behaviors
- team performance classifiers and predictors – team proficiency measures and successful goal achievement at increasing levels of difficulty for representative operational tasks; specific measures of assessment including time, accuracy, etc.; and context descriptors (conditions of performance)
- team proficiency classifiers and predictors – achievements during operational experiences and relevant measures of sustained successful performance; collective standards of expected performance based on rank, roles, and responsibilities within the team

Given we have completed the identification of candidate features for our ML models, our next step is to describe a detailed data management process for feature engineering, model building, and model evaluation in the next section of this paper.

4 A Data Management Process

As noted above, we are proposing a ML approach to identify candidate features, and build a model to predict team learning, performance, and proficiency. In this section, we describe a data management process (Figure 4) for transforming raw data from multiple sources to create inputs or features for our stacked ensemble model. The pipeline includes a data cleaning process, a feature engineering process, and a model building and evaluation process.

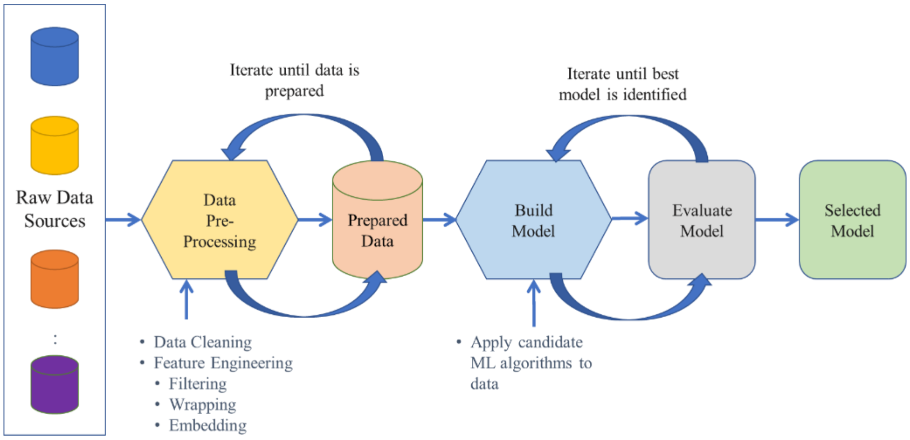

Fig. 4. Data Pipeline Process

The design of the data pipeline assumes multiple streaming sources (most difficult case) instead of a static dataset. Streaming data is also assumed to have both supervised (labeled) and unsupervised (unlabeled) data. To support near real-time processing of data

for the ultimate selected model, our design includes the Massive On-line Analysis (MOA) framework for data stream mining, but our intent is to use this primarily for structuring data flow, data fusion, and data organization. For the deep learning network, we plan to use TensorFlow, a free and open-source software library for ML that can be used across a wide range of tasks but is best for the training and inference of deep learning networks.

During the pre-processing stage, the dataset will be examined and transformed as needed. Features with significant amounts of missing data will be eliminated. Each feature will also be evaluated to ensure good variability since features with poor variability really will not help us distinguish strong associations in the dataset so we can make predictions. Next, we will initiate the feature engineering process. Feature examples for various domains or types of data might include:

- physical attributes – colors, textures, contours, size, mass
- electromagnetic spectrum attributes – frequency, phase, spectrum
- time-based attributes – events, triggers, duration, spacing, trends
- medical domain attributes – DNA sequences, genomes
- textual attributes – words, grammatic dependencies

The success of our ML approach will rely on the available data, the features selected, and model chosen. The goal of the data management process is to find the simplest model that accounts for the major associations in the dataset (reference: Occam's razor). For our team learning, performance, and competency attributes, we expect information sharing, workload sharing, event triggers and other time-based measures to be important features. Our goal in feature engineering is to significantly reduce the number of candidate features to a few critical features for each model. To reduce the number, features may be combined with other features or eliminated based on redundancy or other factors. The benefits of taking time to conduct thorough feature engineering are enhanced generalization of results, faster learning process for your model, improved model interpretability, and improved statistical significance.

The three most used feature engineering methods are filtering, wrapping, and embedding. Filtering strategies look at the interaction of features and eliminate features based on poor correlation between pairs of features or eliminate features based on the statistical significance of individual coefficients in a linear regression model.

Wrapping strategies involve searching through subsets of data, training a model for the current subset evaluating it based on reserved data and then iterating through each subset. One way to approach wrapping is forward selection where the evaluator starts with an empty subset and gradually adds the strongest features. Another is the opposite, backward selection where the evaluator begins with a full set and then removes the weakest feature during each evaluation. Subsets can also be selected randomly and evaluated against each other where the weakest feature in each match is eliminated. Finally, embedded strategies make feature selection part of the model construction process, and use filter and wrapper strategies as part of the model evaluation process. Embedded strategies seek to measure the predictive quality of each subset.

Once the data is prepared and we have selected features for our model, it is time to select ML approaches that are appropriate for our dataset and begin building and evaluating models (Figure 4). Assume we selected to build and test a single DNN (Figure 5) as part of our data management pipeline. DNNs are a class of ML algorithms like artificial neural networks (ANNs) that aim to mimic the information processing of the human brain. DNNs have more than one hidden layer situated between the input and output layers. Generally, accuracy will increase with more hidden layers, but performance will also decrease. However, accuracy is not only dependent on the number of layers, but it is also dependent on the quality of the model, and the quality, characteristics, and quantity of the training data.

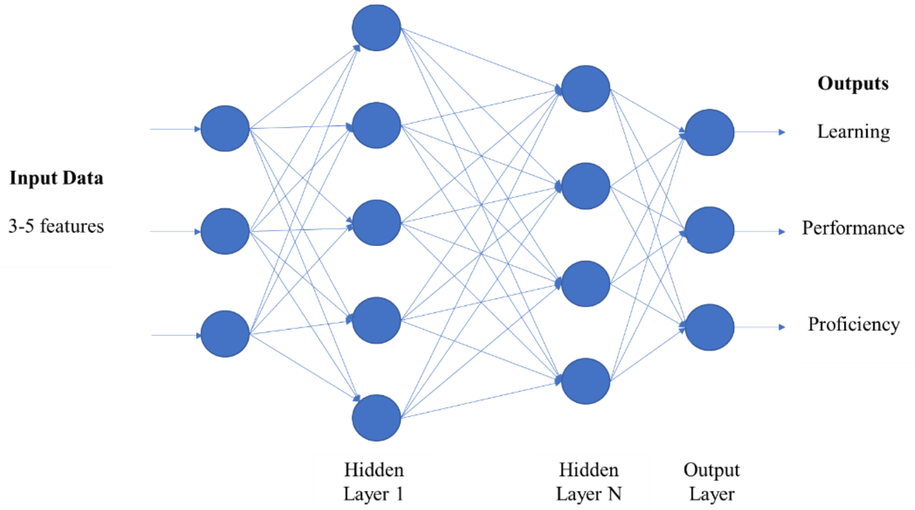

Fig. 5. Deep Neural Network Topography

While a single DNN model may be adequate to assess team learning, performance, and proficiency, a stacked ensemble model is another alternative that should be considered. The rationale for creating a stacked ensemble model is to boost predictive accuracy by combining predictions of multiple, strong, diverse ML models. Each model contributes different strengths, and the individual weaknesses and biases are offset by the strengths of other members. As noted earlier, ensemble methods are commonly used to boost predictive accuracy by combining the predictions of multiple weak ML models to leverage their collective strengths and mitigate their weaknesses. However, a more effective approach is to create an ensemble of strong and diverse models that are stacked to use the power of ML. The simplest kind of ensemble model is the unweighted average of the predictions of the models that form a model library. For example, if a model library includes three strong and diverse models that are used to identify an aircraft maneuver event, each model takes on the same weight when an ensemble model is built.

In a stacked ensemble model, individual model weights are estimated more intelligently by using another layer of ML to optimally combine model outputs and this approach is called model stacking (Figure 6). Model stacking is an efficient ensemble method in which the predictions, generated by using various ML algorithms, are used as inputs in a second-layer learning algorithm. This second-layer algorithm is trained to optimally combine the model predictions to form a new, more accurate set of predictions. For example, a linear regression model may be used as a second-layer model to estimate weights by minimizing the least square errors. Second layer models may also be more complex DNNs or other ML algorithms.

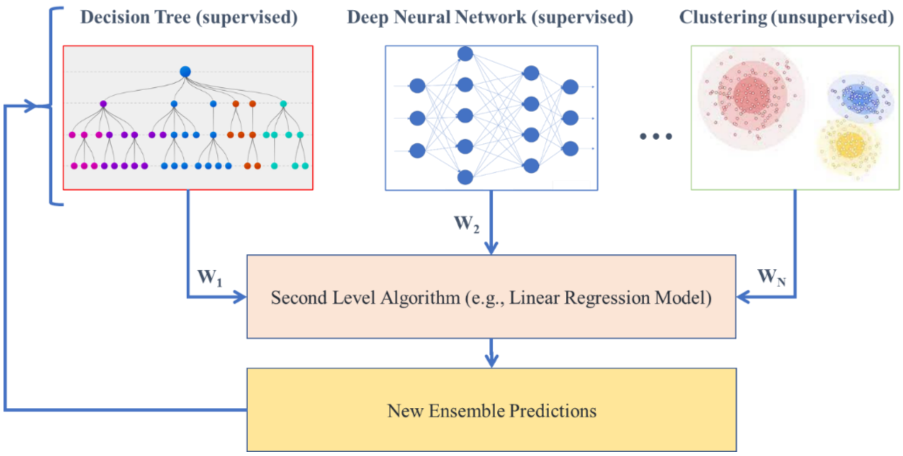
Decision Tree (supervised)
Deep Neural Network (supervised)
Clustering (unsupervised)
W1
W2
WN
Second Level Algorithm (e.g., Linear Regression Model)
New Ensemble Predictions

Fig. 6. Stacked Ensemble Model Development Process

Examining winning entries from the prestigious KDD Cup, an annual knowledge discovery and data mining competition organized by the Association for Computing Machinery (ACM), we found the winning ML designs employed principles that stressed model diversity and were commonly used in building powerful, predictive stacked ensemble models that use:

- ***a variety of training algorithms*** – as many as 64 different models were part of the iterative model evaluation process and these included neural networks, gradient boost models, factorization models, and regression models; an example of applying stacked ensemble modeling methods might be through the addition of a factorization model to a set of tree-based models (such as random forest and gradient boosting) to provide additional diversity which is useful because the factorization model is trained very differently than decision tree models are trained
- ***a variety of features as model inputs*** – seven to ten features were regularly used in different combinations and subsets as inputs for models under evaluation; features could be selected through random sampling of features or through some measure of importance
- ***different data subsets to train models*** – if you have a large dataset, it could be useful to randomly divide the data into subsets or randomly distribute important features across data subsets
- ***different hyperparameter settings*** - hyperparameter are parameters whose values are used to control the model learning process (e.g., model weights); by contrast, the values of other parameters (e.g., accuracy) are derived from the training process; it might be useful to use different hyperparameter settings with different subsets of variables

In the next section, we conclude our discussion by focusing on potential next steps to improve the predictive accuracy of ML solution.

5 Discussion – Next Steps

Most of this paper has focused on a search of the literature to identify candidate features, and a data management process to select features and then build and test ML models to predict team learning, performance, and proficiency. Our exemplar design focused on a deep learning network based on the numerous advantages that it afforded, but we also provide the alternative of a stacked ensemble model with integrated DNN along with other potential ML models (e.g., decision trees, Bayesian networks, clustering algorithms) to ramp up predictive accuracy. We conclude by offering four additional recommendations:

- Employ or develop methods to enrich training repository data (better data = better predictions).
- Extend usable training data with an online ML learning approach to enable team assessments from either static training data sources in a repository or dynamic training data sources (e.g., streaming data from real-time sources).
- Use an online learning approach to process one example at a time and reduce data drift by re-training the ML model after each observation.

- Conduct research to enhance the scalability of ML-based team assessment solutions.

References

1. Wang, M. C., & Walberg, H. J. (1983). Adaptive instruction and classroom time. American Educational Research Journal, 20(4), 601-626.
2. Tsai CC., Hsu CY. (2012) Adaptive Instruction Systems and Learning. In: Seel N.M. (eds) Encyclopedia of the Sciences of Learning. Springer, Boston, MA. https://doi.org/10.1007/978-1-4419-1428-6_1092
3. Sottilare, R., & Brawner, K. (2018, June). Component interaction within the Generalized Intelligent Framework for Tutoring (GIFT) as a model for adaptive instructional system standards. In Proceedings of the 14th International Conference of the Intelligent Tutoring Systems (ITS), Montreal, Quebec, Canada.
4. AIS Consortium Board of Directors (2021). Adaptive Instructional Systems defined. In the Charter of the Adaptive Instructional Systems (AIS) Consortium. Approved 6 January 2021.
5. Biswas, S. (2021). Advantages of Deep Learning, Plus Use Cases and Examples. Published by Width.AI. Retrieved from: https://www.width.ai/post/advantages-of-deep-learning
6. Frensch PA. One concept, multiple meanings: On how to define the concept of implicit learning. Sage Publications, Inc; 1998.
7. Wiliam D. What is assessment for learning? Studies in educational evaluation. 2011 Mar 1;37(1):3-14.
8. Dillenbourg P. Collaborative learning: Cognitive and computational approaches. advances in learning and instruction series. Elsevier Science, Inc., PO Box 945, Madison Square Station, New York, NY 10160-0757; 1999.
9. Salas, E. (2015). Team training essentials: A research-based guide. London: Routledge.
10. Lebas MJ. Performance measurement and performance management. International journal of production economics. 1995 Oct 1;41(1-3):23-35.
11. Grugulis I, Stoyanova D. Skill and performance. British Journal of Industrial Relations. 2011 Sep;49(3):515-36.
12. Davey DD, McGoogan E, Somrak TM, Allen KA, Beccati D, Cramer SF, Frable WJ, Hauser NJ, Hewer EM, Lestadi J, Lulla MK. Competency assessment and proficiency testing. Acta cytologica. 2000 Nov 1;44(6):939-43.
13. Marks, M.A., Mathieu, J.E., & Zaccaro, S.J. (2001). A Temporally Based Framework and Taxonomy of Team Processes. Academy of Management Review, 26, 356-376.
14. Bonner D, Gilbert S, Dorneich MC, Burke S, Walton J, Ray C, Winer E. Taxonomy of teams, team tasks, and tutors. In Generalized intelligent framework for tutoring (GIFT) users symposium (GIFTSym2) 2015 Feb 5 (p. 189).
15. Salas E, Cooke NJ, Rosen MA. On teams, teamwork, and team performance: Discoveries and developments. Human factors. 2008 Jun;50(3):540-7.

Designing Assessments in a Team Trainer for Wargaming

Grace Teo[1], Randy Jensen[2], and Gregory Goodwin[3]

[1]Quantum Improvements Consulting, Orlando, USA
gteo@quantumimprovements.net
[2]Stottler Henke Associates, Inc., San Mateo, USA
jensen@shai.com
[3]U.S. Army Combat Capabilities Development Command – Soldier Center, Orlando, USA
gregory.a.goodwin6.civ@army.mil

Abstract. This paper discusses several concepts for the development of a distributed trainer for command staff trainees learning to develop courses of action (COAs) and wargame. These concepts include how understanding the nature of the team tasks determines the taskwork and teamwork competencies and shapes the pedagogical strategies to be incorporated into the trainer. As well as concepts related to the difficulties in developing assessments for unstructured team tasks and the challenges with assessing team processes, we also discuss the inclination towards a positivist paradigm that relies on the presence of behaviors for indicators, when absence of certain behaviors can also be indicative and used in assessments. We conclude with a preliminary framework for organizing system features for the trainer, and ideas for future research.

Keywords: Taskwork, Teamwork, Pedagogical Strategies, Assessments

1 Introduction

Wargaming is a team activity conducted by command staff with expertise in various warfighting functions (WfFs). The team develops and analyzes courses of action (COAs) by considering critical events, actions, reactions, and counteractions. Their analysis typically results in an operational plan. Any instructional system developed for training wargaming must enable team task execution while also serving training objectives. The extent to which these two goals are met depends on the design of the assessments as well as the training experience and flow of the training exercise delivered by system features. In this paper, we share progress in the design and development of a prototype distributed trainer for U.S. Army wargaming. Our approach to designing assessments for the prototype involves understanding the task and competencies to be trained, and defining indicators or markers of those competencies for assessments. These can only be implemented with system features that can elicit indicative behaviors during task execution and training.

2 The nature of the wargaming team task

The type of team task to be trained should drive many of the system requirements for any trainer. Identifying the type of team task can help researchers and system developers gauge the typical flow of information and communications among team members and anticipate patterns of interactions so that requirements can be designed accordingly. A typology of team tasks had been proposed [1] which describes the ways in which team members work together to accomplish the task. The typology includes characteristics such as complexity, routineness, type of interdependence members have on each other which can be *pooled*, *sequential*, *reciprocal*, or *team* [2]. For instance, members of a team working the assembly line show *sequential* interdependence, while co-authors on a writing team exhibit *reciprocal* interdependence as the writing is passed back and forth between members. Team tasks can also vary in whether they focus on managing, advising, negotiating, performing a service, executing a psychomotor action, or solving problems which can be defined or ill-defined [3]. The wargaming team task can be considered a complex, ill-defined problem-solving task characterized by *team* interdependence where members' tasks and work sequences are unspecified and dynamic. Although it is a challenge to derive any interaction structure for such an unstructured team task with no defined stages requiring a high level of interdependency, there is semblance of some turn-taking in the wargaming team task. For instance, in Division level wargaming, the Chief of Staff typically directs the discussion of COA events and phases, and the staff representing the Intel WfF tends to be called on first to provide background information with intelligence about the enemy. From there, other staff members perform their specialized WfF roles by contributing critical function-specific information as the team steps through various possibilities and topics in their COA discussion. Each team member must not only be familiar with how the different WfFs work together both in planning and execution, but have a nuanced understanding of the COA scenario, including possible repercussions of hypothesized events, and how different WfFs impact further COA decisions down the line.

3 Competencies and pedagogical strategies

An application that enables the command staff team to prepare for and conduct wargaming may not sufficiently support training if it does not support the acquisition of the skills and competencies needed for the task. The team's training must include taskwork and teamwork, and incorporate assessments of outcomes and processes.

3.1 Taskwork and teamwork

While taskwork pertains to *what* the team members do to achieve the collective goal, teamwork focuses on *how* they interact and collaborate to accomplish the team task. A trainer for wargaming should facilitate both taskwork and teamwork in a way that is appropriate for each team member according to their role [4]. Teams that consistently develop good COAs and subsequently wargame effectively would have mastered the needed taskwork and teamwork competencies. There are different levels of taskwork that the trainer should assess; scenario-specific responses such as selecting a "route along the coast which bypasses a mountainous range", and concepts or principles that drive the scenario-specific response, which include assumptions and presuppositions held knowingly or unknowingly by the trainee. These principles are more difficult to assess since they are rarely evident from superficial responses, and may require extra probing before they are apparent. For instance, a further prompt such as "what is the rationale for this selection?" may reveal that the coastal route was selected mainly for speed rather than to minimize risk. If a safer but equally fast route was available, then the coastal route would have been suboptimal for meeting the stated purpose. A trainee may arrive at a suitable scenario-specific response as a mistaken application of a rationale or principle. In wargaming training, it is more crucial to uncover these errors in decision rationale rather than to look for a nominally correct scenario-specific answer, especially for tactical decisions where many options may be acceptable. As these principles are scenario-agnostic, there is a possibility of developing such prompts which can be applied across different scenarios. Instructors and Subject Matter Experts (SMEs) in wargaming training have developed a "mental model matrix" to capture some of these scenario-agnostic principles and topics to address.

The matrix lays out the critical topical areas staff teams must consider for any COA, with a breakdown by WfF. Whether and how trainees address these topics may reveal certain patterns of thinking, including cognitive biases. For example, named areas of interest (NAIs) are intelligence collection points associated with specific locations, typically planned by an Intelligence lead. In explaining how NAI placements contribute to maintaining contact with the enemy, an Intel staff member may realize that they tend to seek out confirmatory evidence. An exercise that involves prompting trainees in this manner can help provide the training experience that promotes development of the right mental models and metacognitive processes.

While instructors and SMEs seek to develop such critical thinking skills and metacognitive awareness in their trainees, they also recognize that the group dynamics within the team can substantially impact how the knowledge and skills of individual WfF staff roles are manifested, drawn out, and sharpened within the team. These interactions constitute the teamwork aspect or process by which the team performs the task. Wargaming assessments are often based on outcome measures such as the synchronization matrix showing the final mission plan and details of dependencies, or the quality of the COAs developed and selected based on criteria such as feasibility, suitability, and completeness. While these may reflect the quality of the team's process, they are not direct measures of it. Training assessments can only provide insight into the team's teamwork if the trainer captures indicators of the team's process, which in turn requires the team competencies to first be defined.

In our approach to ensure that the trainer supports the teamwork needed in wargaming, we identified the team competencies most relevant to COA development and wargaming analysis. These were leadership, team cognition, information exchange, communication quality, supporting behaviors, and team orientation. Teamwork during wargaming involves guiding and directing (leadership), cooperating and offering support (supporting behaviors), and sharing information with each other (information exchange) that builds a shared mental model (team cognition) in a clear and appropriate manner (communication quality) which shows trust and openness (team orientation).

In training for these taskwork and teamwork competencies, instructors and SMEs emphasize the importance of having trainees master the roles of the various WfFs and understand how these must work closely together. They typically adopt the "Socratic" method where trainees are guided through shared dialogs to pose questions, evaluate the reliability of incoming information, cross-examine assumptions by generating alternative explanations or seeking disconfirming evidence [5]. This type of pedagogical approach does not necessarily require high fidelity simulations or even lengthy and complex scenarios. It can be supported by a relatively open system architecture that includes a clear depiction of the vignette of interest, and allows members to respond to open-ended prompts, raise objections and questions, work together to identify decision points, and learn from and build on each other's contributions. From our analysis of the tasks and of instructional methods, we propose that potential pedagogical strategies that support the acquisition of the taskwork and teamwork competencies for wargaming include the following (see Table 1):

Table 1. Summary of wargaming competencies and suggested pedagogical strategy

	Competencies	**Pedagogical/instructional strategy for acquiring competencies**
Taskwork	Individual's knowledge of WfFs, the wargaming process and the military decision making process (MDMP), military protocols and conventions, critical thinking skills, ability to perform assigned staff role	Application of various knowledge and skills to a wide range of scenarios/vignettes and echelons.
Teamwork	Leadership in the team, team cognition, information exchange, team orientation, supporting behaviors, communication quality	Group discussions and questions that reveal and clarify preconceptions, generate ideas to test hypotheses, identify decision points. Active listening practice. Cross-training on staff roles.

It is possible that novice command teams would exhibit different patterns of interactions from expert teams, as will teams composed of members from different services. For instance, all-Army, all-Navy, or mixed-services which will be prevalent in multi-domain operations (MDO), may show different patterns of collaboration due to differences in their tactics, techniques, and procedures (TTPs). Artificial intelligence (AI) can help extract features in team interactions that characterize how different types of teams wargame various types of vignettes and scenarios such as MDOs, operations across echelons, etc.

4 Pre-defining indicators for assessments

In designing assessments for our trainer prototype, we drew from concepts in the Event-Based Approach to Training methodology (EBAT)[6, 7], and the Event Analysis of Systemic Teamwork methodology (EAST)[8, 9]. These methodologies have been successfully applied to training command and control teamwork in aviation and military domains [10–13]. The EBAT involves systematically identifying and injecting events in the training exercise to elicit pre-defined opportunities for observing behaviors indicative of constructs of interest and training objectives. This encourages traceability from behavioral indicators to assessments, and training objectives [14]. The EAST methodology proposes making explicit (i) who the members in the exercise are, (ii) when tasks are performed and who they involve, (iii) where members are, (iv) how members collaborate and communicate to achieve task goals, and (v) what tasks members are performing, and what knowledge and information is shared and used [12]. Both the EBAT and EAST advocate explicitly for articulating and pre-defining anticipated behaviors and markers that serve as measures for constructs of interest. Whereas an event in the EBAT and EAST in an example aviation task may be “reaching cruising altitude” or “initiating the landing procedure”, events in dialog-based, collaborative wargaming preparation tasks could be the prompted discussion of a COA decision point such as anticipating the impact of enemy reinforcements on the scheme of maneuver. Assessment measures for this kind of event concern how well the team discussion covered relevant tactical, cross-functional considerations.

To some extent, such an approach implies a positivist research paradigm which emphasizes positive observations or the presence of behavioral evidence, although the absence of behaviors can also be indicative and should be included in assessments. However, this requires specifying a priori expectations, which are challenging for tasks that are unstructured. For instance, in a well-defined and structured task such as a maintenance task, we can assess absence of certain desirable behaviors when they are not observed in the procedural steps. In unstructured tasks, there are fewer expectations of when certain behaviors should be exhibited, so it is more difficult to note when these are absent. Table 2 presents examples of “positive” (presence of behaviors) and “negative” (absence of behaviors) observations that can be indicators of the teamwork dimensions identified previously. Some are contextualized to wargaming training in a primarily dialog-based exercise environment, and some are more general purpose.

Table 2. Examples of indicators from teams high and low on each team dimension

Team Dimension	Presence of these behaviors	Absence of these behaviors
Leadership (guidance, direction, coordination, strategy formulation)	*High on dimension* -Leader guides who should be doing what, and when. Active team member participation.	*High on dimension* -Absence of team behaviors indicating boredom or distraction
	Low on dimension -Frequent questions on where team is at within the exercise, staff looking at the wrong information	*Low on dimension* -Absence of leader probing questions or indicators of active team engagement
Team cognition (knowing who knows/needs what and the WfF roles, critical thinking,	*High on dimension* -Addresses the right role for information and questions -Discussion appropriate for the echelon -Shows understanding of interdependencies among WfFs, (e.g.,	*High on dimension* -Infrequent requests for clarification on staff roles

Team Dimension	Presence of these behaviors	Absence of these behaviors
metacognitive awareness)	“Signal’s input is needed about maintaining comms, if Aviation takes this helicopter route”) -Questions assumptions, generates alternative explanations, detects missing information, seeks disconfirming evidence	
Team cognition *(continued)*	*Low on dimension* -Addresses the wrong WfF role for questions and information -Jumps to conclusions, reaches for easy explanations, too focused on a quick resolution even with new information	*Low on dimension* -Members not seeking disconfirming evidence, not seeking or using information from other roles
Information Exchange (knowing what info. is needed, how much detail is needed, when to give it)	*High on dimension* -Volunteers information to the right role in a timely manner, with enough detail for the echelon -Discusses topics relevant to the echelon	*High on dimension* -Absence of excessive requests for information (“pulling information”) from staff
	Low on dimension -Gives incomplete, untimely, or inaccurate information -Gives information for the wrong echelon	*Low on dimension* -Absence of appropriate questions, not discussing relevant topics
Supporting Behaviors (backup behaviors, load-leveling, mutual performance monitoring, giving feedback)	*High on dimension* -Shares information that assists others in their work or role -Reminds team of important information missed or overlooked	*High on dimension* -Absence of distracting questions or comments
	Low on dimension -Adds to others’ work even when they are busy -Asks unnecessary questions that distract from topic at hand	*Low on dimension* -Failure to assist or ignores others when they need help
Team Orientation (promotes and supports open communication that facilitates mutual trust, team cohesion, team motivation, conflict resolution)	*High on dimension* -Responsive to each other, shows support (e.g., *like* button) -Encourages contributions, uses “we” often, shares credit/blame as a team -De-escalates and resolves conflicts	*High on dimension* -Not dismissive of others -Does not fuel conflicts
	Low on dimension -Quick to claim credit, shifts blame -Defensive when questioned, pushes own ideas.	*Low on dimension* -Fails to attend to or acknowledge others when they contribute -Lack of participation
Communication Quality (use of proper phraseology, awareness of military conventions)	*High on dimension* -Uses standard conventions and protocols that facilitate clear communications -Communicates concisely -Adjusts communication style as needed	*High on dimension* -Absence of unnecessary chatter
	Low on dimension -Uses wrong terminology that may cause confusion -Uses wrong channels, gives information	*Low on dimension* -Fails to acknowledge others, no closed-loop communications

Team Dimension	Presence of these behaviors	Absence of these behaviors
	in a roundabout way. -Engages in unnecessary chatter	

These measures can be further expanded with more data and research with AI and machine learning (ML). For instance, AI-based methods may detect changes in the frequency and type of questions discussed by a team as they coalesce and work better together. Novel assessment measures may be derived from applying AI and ML approaches to extract emotion indicators from facial expressions, gestures, eye tracking, or vocal data. AI-based speech recognition methods can potentially assist in automating assessments as well.

5 System features for a wargaming trainer

It is relatively straightforward to assess the team's taskwork from responses and outcome measures such as ratings of the tactical decisions within COAs based on criteria set out in rubrics. To assess a team's processes and teamwork, the wargaming trainer must collect data of the members' interactions. These interactions can be scripted into the workflow or be unscripted and ad hoc. Scripted interactions can be embedded in the workflow of the exercise if the team members are required to respond to other members' inputs. These interactions offer the best opportunities for automated assessments. Ad hoc interactions are extemporaneous and can be initiated by any member at any time throughout the team activity. The exercise flow should be designed to create opportunities for both types of interactions to elicit a wide range of behaviors, some of which are indicators of the team competencies. Although the technology in the trainer can be leveraged to automate assessments, assessment opportunities for ad hoc interactions require observer-based assessments. Regardless of how much the interactions are scripted, they are initiated by members receiving information and involve them responding and taking action after processing the stimuli. Given this inevitability, we propose the following framework for system features for the team trainer (see Table 3). Such a framework can serve as a guideline for system development.

Table 3. System features for wargaming taskwork and teamwork

	Features to deliver stimuli	Features that accept inputs
Taskwork	-Display to allow team members to obtain information needed for their roles. *E.g., Intel staff needs information on Commander's critical information requirements (CCIR), and priority intelligence requirements (PIR)* -List of topics to cover in discussion and prompts. *E.g., "What do you need to consider for this situation and what roles are involved?"*	-Interface for submitting products of taskwork, or for assigning taskwork. *E.g., drop-down menu of possible responses and auto-complete options, text box for free form text*
Teamwork	-Display to promote shared understanding across roles. *E.g., common map of area of interest (AOI)* -Display needs to promote awareness of others' status. *E.g., status board showing current status of all staff members*	-Interface to allow initiation of action directed at other members or responding. *E.g., drop-down menu of possible responses and auto-complete options, text box for free form text, messaging with select member(s), "like" button to endorse or acknowledge*

6 Future work and conclusion

Building a trainer to prepare command staff for collaborative wargaming requires attention to both task execution and achieving learning objectives. As with most team tasks, there are taskwork and teamwork competencies to be developed in wargaming training. The challenge lies in the fact that this team task involves few tangible artifacts and does not readily adhere to any type of structure. Being in a cognitive domain of learning [14, 15], the training focus is on the abstract and conceptual, and is difficult to operationalize and assess. In addressing the concepts discussed in the paper, we identified opportunities to apply AI in wargaming training that include learning how types of teams (e.g., novice, expert, mixed services teams, all-Army teams) interact and communicate, exploring if there are patterns of interactions that can be extracted for certain types of vignettes (e.g., operations for different echelons, multi-domain operations), training speech-recognition for the wargaming domain, and developing assessments from novel measures such as eyetracking and gesture and facial expression recognition. We hope to use the prototype currently under development to collect data that can help refine some of these research opportunities.

Acknowledgments. The research reported in this document was performed in connection with contract number W912CG-20- C-0014 with the U.S. Army. The views presented in this paper are those of the authors and should not be interpreted as presenting official positions, either expressed or implied, of the U.S. Government.

References

1. Bonner, D., Gilbert, S., Dorneich, M. C., Burke, S., Walton, J., Ray, C., & Winer, E. (2015). Taxonomy of Teams, Team Tasks, and Tutors, 9.
2. Saavedra, R., Earley, P. C., & Van Dyne, L. (1993). Complex interdependence in task-performing groups. *Journal of applied psychology*, *78*(1), 61.
3. Wildman, J. L., Thayer, A. L., Rosen, M. A., Salas, E., Mathieu, J. E., & Rayne, S. R. (2012). Task types and team-level attributes: Synthesis of team classification literature. *Human Resource Development Review*, *11*(1), 97–129.
4. Crawford, E. R., & Lepine, J. A. (2013). A Configural Theory of Team Processes: Accounting for the Structure of Taskwork and Teamwork. *Academy of Management Review*, *38*(1), 32–48. https://doi.org/10.5465/amr.2011.0206
5. Delic, H., & Bećirović, S. (2016). Socratic Method as an Approach to Teaching. *European Researcher*, *111*, 511–517. https://doi.org/10.13187/er.2016.111.511
6. Fowlkes, J., Dwyer, D. J., Oser, R. L., & Salas, E. (1998). Event-Based Approach to Training (EBAT). *The International Journal of Aviation Psychology*, *8*(3), 209–221.
7. Rosen, M. A., Salas, E., Wu, T. S., Silvestri, S., Lazzara, E. H., Lyons, R., … King, H. B. (2008). Promoting teamwork: An event-based approach to simulation-based teamwork training for emergency medicine residents. *Academic Emergency Medicine*, *15*(11), 1190–1198. https://doi.org/10.1111/j.1553-2712.2008.00180.x
8. Stanton, N., Baber, C., & Harris, D. (2008). *Modelling Command and Control: Event Analysis of Systemic Teamwork*. Ashgate Publishing, Ltd.
9. Walker, G. H., Gibson, H., Stanton, N. A., Baber, C., Salmon, P., & Green, D. (2006). Event analysis of systemic teamwork (EAST): a novel integration of ergonomics methods to analyse C4i activity. *Ergonomics*, *49*(12–13), 1345–1369.
10. Fowlkes, J. E., Lane, N. E., Salas, E., Franz, T., & Oser, R. (1984). Improving the Measurement of Team Performance: The TARGETs Methodology, 16.
11. Harris, D., & Stanton, N. A. (2010). Aviation as a system of systems: Preface to the special issue of human factors in aviation. Taylor & Francis.
12. Salmon, P. M., Lenne, M. G., Walker, G. H., Stanton, N. A., & Filtness, A. (2014). Using the Event Analysis of Systemic Teamwork (EAST) to explore conflicts between different road user groups when making right hand turns at urban intersections. *Ergonomics*, *57*(11), 1628–1642.
13. Smith-Jentsch, K. A., Zeisig, R. L., Acton, B., & McPherson, J. A. (1998). Team dimensional training: A strategy for guided team self-correction.
14. Gagné, R. M. (1972). Domains of learning. *Interchange*, *3*(1), 1–8.
15. Hoque, M. E. (2016). Three domains of learning: Cognitive, affective and psychomotor. *The Journal of EFL Education and Research*, *2*(2), 45–52.