Compound AI Ecosystem: Agents and Tools to Improve Training and Learning

Svitlana Volkova, Summer Rebensky, Laura Cassani, Bob McCormack, Adam Fouse, Sylvain Bruni, Gabe Ganberg and Kara Orvis
Aptima, Inc.
Woburn, MA
{svolkova, srebensky, lcassani, rmccormack, afouse, sbruni, ganberg, korvis}@aptima.com

ABSTRACT

Military training strives to maximize warfighter readiness for mission effectiveness while minimizing the time spent in training exercises. Optimizing training outcomes requires an ability to tailor training to individual learners, considering their previous training, their operational experience, and their role. Compound AI systems, with multiple interacting AI models and tools, offer a promising solution to this problem by combining the advances seen in leveraging large language models (LLMs), multimodal foundation models (MFMs), human digital twins, and multi-agent simulation. This paper describes a future concept in which an AI-driven ecosystem of agents could enhance curriculum-based training. Specifically, we propose a compound AI system and describe a framework for future learning systems that integrates three core components: real-world human training, simulated training, and feedback mechanisms for continuously improving training outcomes. Human learners interact with specialized LLM-driven agents acting as instructors, and separate AI agents to evaluate trainees' competencies, identify knowledge gaps, and generate personalized training content to address deficiencies. To guide real-world instruction, the system leverages simulation using human learner digital twins that model personalities, backgrounds, competencies, and cognitive states. AI agents interact with the digital twins, evaluating their skills and tailoring instructions, generating data to refine instructional tactics, techniques, and procedures for human learners. Ongoing training improvement stems from two additional capabilities: competency evaluation that employs agents to assess human and digital twin proficiencies, and procedural generation of curriculum-relevant content to assist learners struggling with knowledge components. Together, these mechanisms continually optimize instructions to maximize training effectiveness. This novel compound ecosystem of AI-driven agents, simulated human digital twins and their optimized interactions aims to enhance knowledge acquisition, evaluate conditional understanding, and improve the comprehension of unique military concepts. By learning from human-AI exchanges and simulated outcomes, compound AI systems offer a path towards revolutionizing military training.

ABOUT THE AUTHORS

Dr. Svitlana Volkova Chief of AI at Aptima, leads the company's efforts in developing trustworthy and human-centric AI systems that address complex real-world challenges for the Department of Defense and other government agencies. Her research advances natural language processing and machine learning techniques, with a focus on graph neural networks, causal inference, and multimodal models. Dr. Volkova's pioneering work in AI-powered analytics has made significant contributions to explaining complex social systems and behaviors. A recognized leader in human-centered AI design, evaluation, and trustworthy AI systems, Dr. Volkova has spearheaded projects funded by DARPA, IARPA, SOCOM, DOE ASCR, and NNSA, advancing AI capabilities to support critical national security missions. Her expertise is reflected in over 70 peer-reviewed publications spanning AI, machine learning, and social media analytics. Dr. Volkova serves on program committees and review boards for top-tier AI conferences, including AAAI, ACL, EMNLP, and NeurIPS. As an advocate for diversity in technology, she is an active member of Women in Machine Learning. Dr. Volkova earned her PhD in Computer Science from Johns Hopkins University, where she was affiliated with the Center for Language and Speech Processing and the Human Language Technology Center of Excellence. Her forward-thinking approach and commitment to responsible AI development position her at the forefront of shaping the future of safe, secure, and trustworthy AI that align with human values and societal needs.

Dr. Summer Rebensky Scientist, Aptima, Inc. has expertise focusing on human performance, cognition, and training in emerging systems. In her role at Aptima, Inc. she serves as the capability lead for Air Force Training, Learning, and Readiness technologies. She specializes in leading efforts to develop, test, and implement AR, VR, XR, and modern training solutions to improve and measure human performance. Dr. Rebensky has previous experience as a research fellow as a part of the Air Force Research Laboratory's Gaming Research Integration for Learning Laboratory

(GRILL) conducting research on interface designs for novel aviation use cases in the civilian and DoD world as well as and human-agent teaming designs. Her other experience includes leading Florida Tech's ATLAS research lab with grant and STTR research with the Air Force, Navy, and FAA; human factors assessments of a trainer aircraft for the F-35 with Northrop Grumman; and developing DoD courseware with Raytheon. Her research experience involves leveraging VR, AR, and modern technology to optimize human performance in training and operations, individualizing learning through gamification and adaptive technologies, human-agent teaming, and trust in AI. Dr. Rebensky received her BA in psychology, MS in aviation human factors, and PhD in aviation sciences with a focus in human factors from Florida Tech.

Ms. Laura Cassani, Principal Research Engineer, Aptima, Inc. is also the Deputy Director of the Intelligent Performance Analytics (IPA) Division, which focuses on innovation in the development and applications of artificial intelligence technologies to improve human and machine performance. Previously, Ms. Cassani led the System Optimization Analytics (SOA) Capability, a portfolio of projects focused on operationalizing artificial intelligence approaches to dynamic human-machine systems. She has led several successful research and development efforts for the DoD, serving as Principal Investigator on multiple projects for ONR, DARPA, AFRL, and MCSC. Ms. Cassani has focused most recently on leading projects involving large language models (LLMs) and generative artificial intelligence. Ms. Cassani serves as the Principal Investigator for the test and evaluation technical area for DARPA's Semantic Forensics (SemaFor) program, which focuses on developing measurement methods to evaluate algorithms that aim to detect, attribute, and characterize semantic inconsistencies in multimedia and synthetically generated audio, video, text, and image data. Ms. Cassani also managed the DARPA Civil Sanctuary program that built an information operations testbed utilizing trained generative AI personas for experimentation. For the past 12 years, she led technology development for the Marine Civil Information Management System (MARCIMS), a deployed Program of Record supporting Marine Corps Civil Affairs operators. Ms. Cassani managed engagements with military operators and key stakeholders through joint exercise and experimentation both OCONUS and through collaboration in developing training materials with stateside entities including the Marine Corps Civil Military Operations School. Ms. Cassani holds a MA in security studies from Georgetown University and a BA in international relations from Boston University.

Dr. Robert McCormack, Principal Research Engineer and Director of the Intelligent Performance Analytics Division, Aptima, Inc. has spent more than15 years working with scientists and engineers to develop and deliver data-driven solutions for understanding and predicting individual and team performance dynamics. With expertise in natural language processing, machine learning, and statistical methods, he has extensive experience analyzing both structured and unstructured data to extract meaningful information. Dr. McCormack has led a variety of efforts while at Aptima. Most recently, he has helped to develop the Teams Research Assessment Kit (TRAK), an integrated hardware and software suite that utilizes unobtrusive sensors to capture and analyze the real-time dynamics of team performance. Dr. McCormack also has served as lead engineer on PARSERS, an effort aimed at identifying suicide risk in a civilian workforce that utilizes explainable machine-learning techniques. Dr. McCormack has served as the Principal Investigator on more than a dozen projects at Aptima. He received a PhD and MS in mathematics from Texas Tech University, and a BA in mathematics and computer science from Austin College.

Mr. Sylvain Bruni, Principal Engineer and Deputy Division Director for Performance Augmentation Systems, Aptima, Inc. His technical work focuses on developing and deploying human-AI collaborative systems and technologies in defense and healthcare domains. Mr. Bruni supervises a portfolio of defense and civilian cognitive assistants called Sidekick™ (Systems for Interactive Discovery and Exploitation of Knowledge and Insights with Contextual Kinetics). At Aptima, he provides expertise and leadership in human factors and cognitive systems engineering, design thinking, rapid prototyping, and agile methodologies (including DevOps and SAFe). In 2022, Mr. Bruni became an Entrepreneur-in-Residence at Aptima, with a goal of transitioning two company applications to commercial markets: a Sidekick™ for remote patient monitoring, and a caregiver engagement platform in neonatal intensive care units and other critical care settings. Mr. Bruni is the current Co-Chair of Aptima's Inclusion, Diversity, Equity, and Antiracism (IDEA) Committee, and the Chair of the Council of Affinity Groups (COAG) at the Human Factors and Ergonomics Society (HFES), following two years as founding Co-Chair of its LGBTQ+ Affinity Group. Mr. Bruni holds an SM in Aeronautics and Astronautics from Massachusetts Institute of Technology (MIT) and a Diplôme d'Ingénieur from CentraleSupélec (France). He is a trained SAFe Advanced Scrum Master and Product Owner and a certified CMMI v2 Associate. Mr. Bruni is a member of HFESand the INCOSE, ACM, , SMC, IEEE Human Systems Integration (HSI) Standards Committee of SAE.

Dr. Adam Fouse, Principal Research Engineer and Senior Division Director, Performance Augmentation Systems Division, Aptima, Inc. leads a team of researchers, designers, and engineers in creating tools to increase human performance in areas such as information analysis and industrial health and safety. In addition to his division leadership, Dr. Fouse has served as a principal investigator on numerous research projects for customers across the DoD, including Aptima's efforts on the DARPA A-Teams and ASIST programs. His research applies an interdisciplinary approach that combines cognitive, social, and computational sciences to address a diverse set of projects ranging from context-aware systems for information analysts to computational modeling of human-AI teams. Dr. Fouse holds a PhD and MS in cognitive science from the University of California, San Diego, and a BA in cognitive science and computer science from Brown University. He is a member of the Association for Computing Machinery and the Human Factors and Ergonomics Society.

Mr. Gabriel Ganberg, Chief Architect, Aptima, Inc. has more than 20 years of experience leading and architecting software projects in the R&D space. Serving as the Chief Architect for Aptima's Research & Engineering group, his focus is on building common platforms that support the cutting-edge R&D programs happening across the company. Mr. Ganberg architected and serves as technical lead for multiple Aptima platform technologies including AI Toolbox (LLMs, analytics, RAG, etc.), CRAFT (long form document and multi-dimensional analysis generation through LLMs), Discourse (generative testbed for conversation analysis and simulation), HAT (human-AI teaming), and YAADA (data-engineering infrastructure and analytic framework). Mr. Ganberg builds software in a variety of domains, including intelligence analysis, business intelligence, cyber analytics, air traffic control, human/AI collaboration, recommendation systems, and experimental team research. He develops tech stacks that leverage cloud architecture, distributed systems, document/graph/vector databases, DevOps, and applied AI/ML. Mr. Ganberg received a BA in computer science and economics from Vassar College.

Dr. Kara Orvis, Executive Vice President, Research and Engineering, Aptima, Inc., is also a principal scientist with more than 20 years of experience in government research and development across the military services. Her expertise is in the areas of training, leadership, teams, culture, distributed work, and unobtrusive measurement for which she has more than 70 publications/ presentations, including one edited book. At Aptima, she leads projects related to military assessment, formation, training, and development. As an Army contractor, she has successfully supported over 12 contracts for ARI units at Ft. Hood, Aberdeen, Orlando, Ft. Leavenworth, and Headquarters, serving as PI for the majority of those. Over the past decade, Dr. Orvis has dedicated her research to understanding the use and value of data to help solve assessment and personnel issues. She served as PI for the project that proceeded this proposal. She also served as PI for the ACCRUE project in which unobtrusive measures were validated to assess Army staff processes. She also led the development of a tool, TeamBuilder, to staff teams to job descriptions using language analytics and developed a tool to map job experiences to the development of Army leadership competencies (LeadershipMAP). Dr. Orvis has been awarded two patents related to using text-based data to assess teamwork skills and to match teams to job descriptions and is currently working on a provisional patent specifying a process by which machine learning approaches can be matched to cognitive work tasks in support of developing AI solutions. Dr. Orvis holds an MA and a PhD in industrial-organizational psychology from George Mason University and a BA in psychology from Ohio Wesleyan University. She is a member of the American Psychological Association and the Society for Industrial and Organizational Psychology.

Compound AI Ecosystem: Agents and Tools to Improve Training and Learning

Svitlana Volkova, Summer Rebensky, Laura Cassani, Bob McCormack, Adam Fouse,
Sylvain Bruni, Gabe Ganberg and Kara Orvis
Aptima, Inc.
Woburn, MA
{svolkova, srebensky, lcassani, rmccormack, afouse, sbruni, ganberg, korvis}@aptima.com

INTRODUCTION

The effectiveness of military training is critical for ensuring the readiness and mission success of warfighters. As stated by the U.S. Department of Defense, "the primary purpose of military training is to prepare forces for combat and to ensure that they can accomplish their assigned missions" (Department of Defense, 2020). Optimizing training outcomes is crucial for preparing military personnel to face complex and evolving challenges in an increasingly dynamic and uncertain global security environment (Joint Chiefs of Staff, 2018).

However, optimizing training outcomes poses significant challenges. Tailoring training to individual learners, considering their previous experiences, operational backgrounds, and roles, is a complex task (Sottilare et al., 2018). Traditional training methods often struggle to provide the level of adaptability and personalization required to maximize the effectiveness of military training (Fletcher, 2009). Moreover, the rapid advancement of technology and the changing nature of warfare necessitate the development of innovative training solutions that can keep pace with these changes (NATO Science and Technology Organization, 2021).

Compound AI systems, which combine multiple interacting AI models and tools, offer a promising solution to address these challenges. As Zaharia et al. (2024) discuss, compound AI systems tackle tasks using multiple interacting components, including multiple calls to models, retrievers, or external tools, thus have the potential to revolutionize training and learning. LLMs, such as GPT-4 and Anthropic's Claude, have demonstrated remarkable capabilities in natural language processing, problem-solving, and knowledge generation (Brown et al., 2020; Anthropic, 2023). Multimodal foundation models, such as DALL-E 2, Imagen, and Flamingo, have showcased impressive abilities in generating, manipulating, and understanding visual and textual data, enabling the creation of highly realistic images and videos from natural language descriptions (Ramesh et al., 2022; Saharia et al., 2022; Alayrac et al., 2022). AI-driven agents and tools, such as virtual assistants, chatbots, and recommendation systems, have demonstrated significant potential in enhancing user experiences, automating tasks, and supporting decision-making processes across various domains, including education, healthcare, etc. (Gao et al., 2018; Srivastava et al., 2019; Lu et al., 2020). Many service members speak openly about beginning to use these AI models and tools to update content in their classrooms and automate some of their more manual tasks. However, each AI system has its own limitations and best use cases. We posit a compound-AI system, that leverages the best of each system to deliver the best training.

By leveraging these cutting-edge technologies in a novel compound AI system, we can significantly enhance the effectiveness and efficiency of training, ultimately leading to improved warfighter readiness and mission success. The proposed approach is innovative in its integration of multiple AI components, each addressing specific aspects of the training process. LLMs and MFMs can generate adaptive content and provide personalized instruction, while digital twins enable realistic simulations tailored to individual learners. Multi-agent simulations allow for the optimization of instructional tactics and strategies through the interaction of AI agents with human learner digital twins. This comprehensive, AI-driven approach to military training is a departure from traditional methods, offering unprecedented levels of adaptability, personalization, and efficiency. By continuously learning from real-world and simulated training data, the compound AI system can refine and optimize its strategies, creating a virtuous cycle of improvement. The potential impact of this transformative approach extends beyond military training, as the principles and technologies employed can be adapted to various domains. This paper will discuss the design of a compound-AI system and where each AI-system can provide the greatest value.

COMPOUND AI SYSTEM OVERVIEW

Integration of real-world human training, simulations, and feedback

The compound AI system presented in this paper envisions a novel combination of technologies – LLMs, MFMs, digital twins, multi-agent systems, specialized memories, and predictive/causal ML – in a cohesive compound AI system that bridges real-world and simulated training that offers immense potential to adaptively personalize instruction, accelerate expertise development, and ultimately transform how we approach training and education across domains. Key aspects of this compound AI system for enhancing training include:

- Integrating multiple specialized AI components, including LLMs and MFMs serving as **instructor assistant agents**, **evaluator agents**, human learner digital twins for simulation, and **reasoning agents** deriving insights from the training data.
- Orchestrating a bi-directional flow between the small-scale real-world training with human learners and the large-scale simulated training environment. Observational data from human training sessions can initiate simulations to optimize instruction, while interventional data from simulations with digital twins can generate curriculum improvements for the real-world training which allows for continual optimization.
- Developing AI-powered instructor and tutor agents can proactively evaluate learner competencies, identify knowledge gaps, and procedurally generate personalized content and curriculum to adaptively address those gaps.
- Building semantic, episodic and declarative memory components enable the AI agents to build and leverage knowledge over time from their interactions and experiences to further optimize training.
- Enabling predictive analytic tools can anticipate challenges and knowledge gaps, allowing preemptive interventions. Causal inference tools can explain the drivers behind competency gaps and prescribe remediations.

Figure 1 demonstrates how combining various AI components like episodic, declarative and semantic memory, experiences, past conversations, and a knowledge base can enable a compound AI system that interacts with and learns from both small-scale real-world human training and large-scale simulated training with digital twins and AI agents.

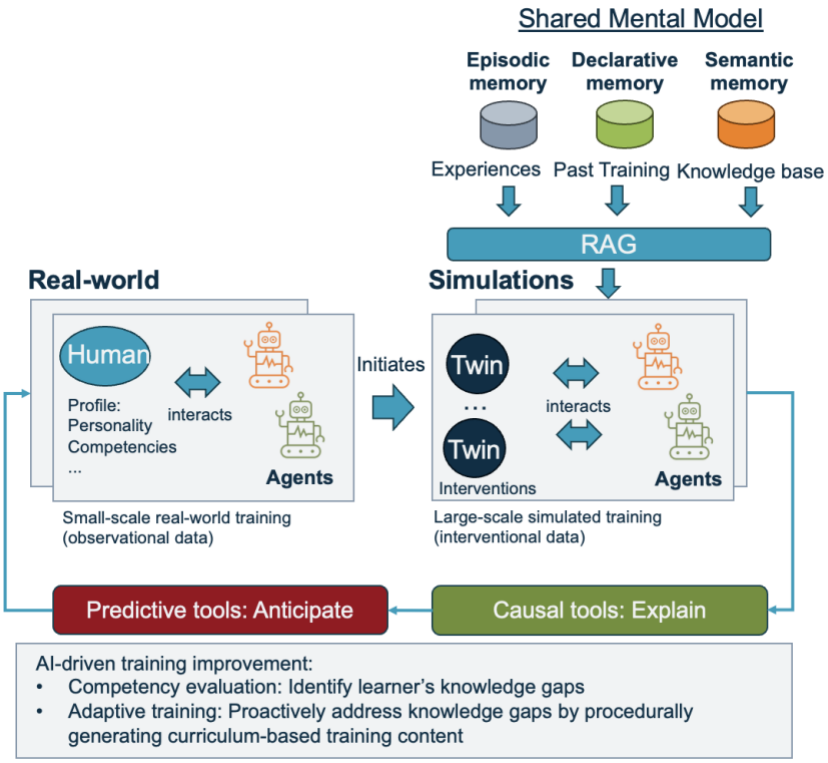

Figure 1. Proposed Compound AI Ecosystem for Training and Learning

The concept of **human digital twins**, which are virtual representations of individual learners, has gained traction in recent years to personalize learning experiences and optimize training outcomes (Madni et al., 2019). By modeling learner personalities, backgrounds, competencies, and cognitive states, digital twins enable the creation of realistic simulations tailored to individual needs (Karakra et al., 2019). This approach allows for the safe and controlled development of skills, as well as the assessment of performance in various scenarios. The compound AI system incorporates human learner digital twins to enable highly realistic and personalized simulated training experiences. These digital twins are data-driven, virtual representations of individual learners, encompassing their unique characteristics, backgrounds, competencies, and cognitive states (Beal and Jaine, 2022; Sammut 2021, Mazhari et al., 2021). By leveraging LLMs and MFMs, the system can create accurate and dynamic models of learners, which evolve in real-time based on their interactions, performance, and learning progress. To create highly realistic and personalized digital twins, the compound AI system leverages advanced modeling techniques to capture various aspects of learners' **personalities, backgrounds, competencies, and cognitive states.** LLMs, with their ability to understand and generate natural language, can process learner profiles, performance histories, and interaction data to infer individual traits, preferences, and learning styles (Brown et al., 2020). MFMs, such as Flamingo, can analyze multimodal data, including facial expressions, vocal cues, and physiological signals, to gauge learners' emotional states and levels of engagement. By integrating these data points, the system can create holistic and dynamic representations of learners, enabling more accurate and adaptive simulated experiences. Digital twins in a training scenario can be used in one of two ways (1) during a learning instance, simulating the experience of the learner in a variety of settings and finding the one with the most positive outcomes, and delivering that training, or (2) exploring novel concepts of training to identify potential long-term impacts. These methods could be implemented within a system such as MyLearning to deliver the next recommended training content or full online training course paths.

Multi-agent simulation, another key component of compound AI systems, involves the interaction of multiple AI agents within a simulated environment (Wooldridge, 2009). In the context of military training, these agents can represent instructors, learners, and various entities within a training scenario. By allowing AI agents to interact with human learner digital twins, the compound AI system can optimize instructional tactics and strategies, ensuring that training is both effective and efficient (Zaharia et al., 2024). In wargaming scenarios, LVC scenarios, or in platforms such as virtual battlespace, multi-agent simulation could provide: (a) simulated red force entities, (b) live summarizations to instructors on performance, (c) inject challenges and learning interventions, and (d) provide immediate performance calculations to lead informed after-action review.

The integration of real-world human training, simulated training, and feedback mechanisms is essential for the continuous improvement of training outcomes (Nielson & Kratiak, 2021). Real-world training provides authentic experiences, while simulated training offers a safe and controlled environment for skill development. The AI-systems can learn and then adapt over time to base upon live data it receives. Feedback mechanisms ensure that insights from both real-world and simulated training are used to refine and optimize the overall training process, creating a virtuous cycle of improvement.

LLM-driven agents enhancing curriculum-based training

The compound AI system should leverage an ecosystem of AI agents powered by large language models (LLMs) to enhance curriculum-based training. These LLMs, such as GPT-4, Anthropic's Claude, Mistral or LLaMa, serve as the foundation for generating adaptive content, providing personalized instruction, and facilitating seamless interaction between AI agents and human learners. The LLMs' deep understanding of natural language and ability to generate human-like responses enable the AI agents to engage in meaningful dialogue, answer questions, and provide explanations tailored to each learner's needs (Brown et al., 2020; Anthropic, 2023).

Specialized LLM agents as instructors, evaluators and reasoners

The compound AI system can also employ specialized LLM agents to serve as instructors, evaluators and reasoners across various training domains. These domain-specific agents possess deep knowledge and expertise in their respective fields, enabling them to provide targeted instruction, answer complex questions, and evaluate learner performance with high accuracy. By leveraging the capabilities of LLMs, these specialized agents can understand the nuances of learner inquiries, provide contextually relevant explanations, and offer guidance tailored to individual learners' backgrounds and skill levels (Gao et al., 2018; Srivastava et al., 2019). For example, **content generation agents** can generate rich, multimodal content, such as text, images, videos, and simulations. Performance **evaluation agents** can analyze learner submissions, such as written assignments, practical exercises, and simulated scenarios, to

evaluate their understanding, skill application, and adherence to best practices. **Reasoner agents** can process and analyze vast amounts of training data, including learner performance metrics, feedback, and operational readiness reports, to identify systemic issues, skill deficiencies, and areas requiring additional focus. Table 1 further defines the example types of agents present in the compound AI ecosystem for training, the intended purpose of the agent, the AI components needed to implement the agent, and operational impact of the agent on training processes and outcomes.

Table 1. Proposed agent types, purpose, impact and corresponding compound AI system components

Example Agent Types	Purpose	Compound AI System Components	Impact
Content Generation Agents	Generate multimodal content, analyze performance, and plan curriculum	LLMs, MFMs e.g., GPT-4, DALL-E 2	Enables instructors to quickly create engaging and interactive learning materials tailored to specific learning objectives and learner needs
Performance Evaluation Agents	Assess learner performance and provide feedback	LLMs and ML tools	Enables scalable, consistent, and objective assessment of learner performance, reducing the burden on instructors and facilitating timely, actionable feedback to support learner growth and development.
Reasoning Agents	Generate insights and situational awareness of training gaps	LLMs and analytics tools	Enabling informed decisions regarding resource allocation, training priorities, and curriculum improvements.

AI Agents & Training Interactions in Practice

The interactions between AI agents and human learner digital twins in the simulated training environment allows for the optimization of instructional tactics and strategies. AI agents, serving as content generators, evaluators and reasoners, can engage in realistic, personalized interactions with digital twins, leveraging their knowledge of individual learner characteristics and performance histories. Possible interventions include adjusting competency ranges to simulate different skill levels, modifying personality traits to assess their impact on learning outcomes, altering past learning experiences to understand how prior knowledge affects new skill acquisition, varying cognitive states to optimize content delivery timing, simulating different operational backgrounds to tailor training for specific roles, and adjusting learning styles to find the most effective instructional methods.

To enhance this optimization process, the system incorporates advanced predictive and causal models. Predictive models, such as those based on machine and deep learning algorithms like random forest (Jordan and Mitchel, 2015), long-short term memory models (Greff et al., 2022; Volkova et al., 2017), transformer architectures (Vaswani et al., 2027; Horawalavithana et al., 2022) or graph neural networks (Zhou et al., 2020; Wu et al., 2020; Shrestha et al., 2019a; Shrestha et al., 2019b; Horawalavithana et al., 2023), can forecast learner performance and identify potential challenges before they arise. For instance, Volkova et al. (2017) demonstrated the effectiveness of using linguistic and behavioral features to predict future outcomes, which can be adapted to anticipate learner progress and potential stumbling blocks in the training process. If we know how a trainee at a current level of competency or proficiency will respond to specific training exercises ahead of the fact, we can simulate the outcomes of various training sessions to identify the one with ideal outcomes. At scale, this can also be done to build entire training syllabi depending on the unique trainee. Predictions can also be run to determine when supplemental training is best delivered, when skill decay will occur, or when currency training would be best administered.

Causal inference tools, on the other hand, help identify the underlying factors driving learning outcomes. Techniques such as causal structure learning (Pearl 2009; Spirtes et al., 2000) or average treatment effect estimation algorithms like causal forest (Wager and Athey, 2018) can be employed to understand the causal relationships between various interventions and their effects on learner performance. Volkova et al. (2023) showcased the use of causal discovery on explaining human social behavior, following prior work by Glenski and Volkova (2021) and Saldanha et. al. (2020), which can be adapted to the training context to determine which specific instructional strategies or environmental factors have the most significant impact on learning outcomes. For example, the impact of moving to condensed learning schedule of four 10-hour days as opposed to five 8-hour days. Furthermore, the system can utilize counterfactual reasoning models to simulate "what-if" scenarios, allowing for the exploration of alternative training approaches without the need for real-world implementation (Guo et al., 2021; Cottam et al., 2021).

By integrating these predictive and causal tools, a compound AI system can anticipate learning curves and adjust difficulty levels proactively, identify the most influential factors in skill acquisition for different learner profiles, optimize the sequencing of training modules based on causal relationships between skills, personalize interventions by understanding the causal impact of various strategies on individual learners, and simulate long-term outcomes of different training approaches using counterfactual models. These advanced analytical capabilities allow the system to move beyond simple correlation-based optimizations and towards a deeper understanding of the causal mechanisms underlying effective training. By continuously learning from the outcomes of these simulated interactions and interventions on the digital twins, and leveraging the insights from predictive and causal models, the compound AI system can adaptively optimize its instructional tactics with greater precision and effectiveness. This leads to improved learning outcomes and accelerated skill development when applied to real-world training scenarios, ultimately enhancing the readiness and capabilities of military personnel in facing complex and evolving challenges.

TRAINING IMPROVEMENT MECHANISMS

The integration of predictive and causal models into the compound AI system offers unprecedented opportunities to enhance military training effectiveness. By leveraging machine and deep learning techniques and causal inference methods, we can create a dynamic, adaptive training environment that continuously optimizes learner outcomes.

I. **Competency evaluation and adaptive training agents and tools.** AI-driven evaluation to identify learner knowledge gaps is a critical component of a compound AI system. Building upon the work of Sottilare et al. (2018) in adaptive instructional systems, we recommend employing ensemble machine learning models that combine multiple data sources, including performance metrics, interaction patterns, and physiological signals. These models can identify subtle patterns indicative of knowledge gaps or misconceptions. For example, we can adapt the linguistic and behavioral feature analysis techniques developed by Volkova et al. (2017) to the training context, enabling real-time assessment of learner comprehension and skill proficiency. Adaptive training to procedurally address gaps leverages these evaluations to dynamically adjust the training curriculum. This approach could extend beyond traditional adaptive learning systems by incorporating causal inference techniques. By employing causal algorithms (Volkova et al., 2023), we can identify the most effective interventions for specific knowledge gaps and learner profiles. This allows for precise, personalized training adjustments that maximize skill acquisition efficiency.

II. **Procedural content generation agents.** A novel approach that combines LLMs with MFMs with a Retrieval Augmented generation (RAG) backend to generate diverse, personalized training materials could provide tailored content for struggling learners. This builds upon the work of Ramesh et al. (2022) and Brown et al. (2020) in generative AI but extends it to the specific context of military training. A generative AI system with the RAG backend can create scenario-based exercises, interactive simulations, and multimedia explanations tailored to individual learner needs and learning styles can greatly reduce time to develop content (Aptima, 2023). Enhancing knowledge acquisition and comprehension is achieved through the integration of cognitive science principles into content generation algorithms.

III. **Performance analysis, predictive and causal tools for real-time insights.** A compound AI system that employs cutting-edge predictive models and tools to provide real-time insights into learner performance and potential outcomes can provide accurate real-time insights. Building upon the work of Volkova et al. (2021) in multi-platform information analysis, adapting their techniques to the training domain. By using counterfactual reasoning models, inspired by the work of Volkova et al. (2023), the models allow simulation of alternative training scenarios and their potential outcomes, enabling proactive optimization of training strategies. For example, we can predict the long-term impact of different intervention strategies on a learner's skill development, allowing instructors to make informed decisions about resource allocation and training focus. Furthermore, we incorporate causal discovery algorithms (Glenski and Volkova, 2021) to uncover the underlying causal structures in the training process. This allows us to move beyond mere correlation and identify the true drivers of training effectiveness. By understanding these causal relationships, we can design more efficient and targeted training interventions.

COMPARISON WITH EXISTING TRAINING SYSTEMS

While traditional military training systems have incorporated elements of simulation and computer-based instruction, a compound AI ecosystem offers several key advancements as shown in Table 2.

Table 2. Compound AI System Tactical Advantages Compared to Traditional Training Methods

Requirement	Compound AI System Tactical Advantages
Personalization	Unlike standard simulation-based training, it provides a level of personalization that adapts not just to performance, but to individual learning styles, cognitive states, and career trajectories.
Scalability	Unlike traditional systems often require significant human oversight it allows for scalable, consistent training experiences across large numbers of trainees simultaneously.
Continuous Improvement	Unlike static training programs, a compound AI system that uses causal inference and digital twins allows for continuous optimization of training strategies based on comprehensive data analysis.
Multimodality	While existing systems might use video or text-based scenarios, the integration of LLMs and MFMs allows for richly detailed, dynamically generated multi-modal training content.
Predictive Capabilities	Unlike traditional systems that are largely reactive the use of predictive analytics allows for proactive identification of potential skill gaps and tailored interventions.

BENEFITS AND FUTURE POTENTIAL

The compound AI ecosystem for enhancing training and learning presented in this paper envisions a significant leap forward in educational technology, particularly for military applications. In this section, we explore the key advantages of the approach and discuss its potential for future development and application across various domains. From maximizing training effectiveness through optimized human-AI interaction to enabling rapid, scalable content development, our compound AI system paves the way for a new era in personalized, adaptive, and highly efficient training solutions.

Maximizing training effectiveness by optimizing human-AI interaction: The compound AI system presented here offers unprecedented opportunities to maximize training effectiveness through optimized human-AI interaction. By leveraging the strengths of both human instructors and AI agents, we can create a synergistic learning environment that adapts in real-time to learner needs. As demonstrated by Fouse et al., (2018), Ezer at al., (2019), Fouse et al., (2019), Born et al., (2023), and Chaparro-Osman et al., (2023) in their work on human-systems integration, such collaborative approaches can lead to significant improvements in performance and decision-making.

Scaling personalized learning through automation: The integration of LLMs and MFMs in the system enables the scaling of personalized instruction to an unprecedented degree. Building of recent work on generative AI, a compound-AI system can generate tailored content, explanations, and assessments for each learner, addressing the long-standing challenge of providing individualized education at scale (Fletcher, 2009). This automation of personalized instruction not only increases the reach of high-quality training but also ensures consistency in content delivery while adapting to individual learner needs.

Continual improvement via human and AI feedback loops: A compound AI system such as this could incorporate continual improvement mechanisms through compounded human and AI feedback loops. This approach builds upon the concept of human-in-the-loop machine learning (Xin et al., 2018) but extends it to create a multi-layered feedback system. Real-world training data informs simulations with digital twins, which in turn generate insights to refine real-world training approaches. This iterative process, combined with the predictive and causal models discussed earlier, creates a self-improving system that becomes more effective over time, adapting to evolving training needs and incorporating new knowledge seamlessly. When implemented within a training program, the machine learning mechanisms would be robust to changes in doctrine, training paradigms, and instructors.

CHALLENGES AND LIMITATIONS

Compound AI ecosystems offer significant potential benefits for military training. However, it is important to acknowledge and address the challenges and limitations associated with implementing such a complex system.

a) *Technical challenges:* The integration of multiple AI components, including LLMs, MFMs, and digital twins, requires substantial computational resources and sophisticated software architecture. Ensuring seamless interaction between these components while maintaining real-time performance presents a significant technical hurdle.
b) *Data privacy and security:* Military training often involves sensitive information and scenarios. Implementing robust data protection measures to safeguard personal information of trainees and classified training content is crucial. The system must comply with strict military security protocols while still maintaining the flexibility to adapt and learn from training data.
c) *AI bias and fairness:* As with any AI system, there is a risk of bias in the training data or algorithms, which could lead to unfair or inaccurate assessments of trainees. Mitigating bias in the digital twin models, content generation, and evaluation processes is essential to ensure equitable training outcomes for all.
d) *Ethical considerations:* The use of AI in military training raises important ethical questions. There are concerns about the potential over-reliance on AI-generated content and assessments, which could impact human decision-making skills and accountability. Furthermore, the use of digital twins and extensive personal data collection for training purposes may raise privacy concerns among trainees.
e) *Human factors and adoption:* The successful implementation of this system depends on its acceptance by both trainers and trainees. Resistance to change, concerns about job displacement, and skepticism about AI-driven training methods may pose challenges to widespread adoption.

To enable the successful implementation of compound AI ecosystems in military training, the military must develop clear policies and regulations governing the use of AI in classified spaces, balancing data privacy, security, and ethical considerations with the necessary flexibility to harness the benefits of AI. Furthermore, the military should allocate resources to support the creation of robust, secure, and scalable AI training platforms, as well as fund research initiatives focused on addressing technical challenges, mitigating bias, and ensuring fairness in AI-driven training systems.

CONCLUSIONS

The compound AI system we have proposed represents a significant leap forward in the field of training and education. By integrating cutting-edge AI technologies such as LLMs, MFMs, digital twins, and multi-agent simulations with advanced predictive and causal models, we have described a framework that could dramatically accelerate expertise development. This envisioned approach goes beyond traditional adaptive learning systems by incorporating causal inference, counterfactual reasoning, and real-time optimization of training strategies. The integration of real-world and simulated training environments creates a robust, data-driven system that continuously improves its effectiveness. As demonstrated by the potential applications across various domains, this compound AI approach has the power to transform how we approach skill development and knowledge acquisition across society.

Looking forward, we envision a future where AI-enabled training systems become an integral part of lifelong learning and skill development. As AI technologies continue to advance, we anticipate even more sophisticated integration of multimodal inputs, including physiological data, environmental factors, and social dynamics, to create holistic learning experiences. The development of more advanced causal models may allow for even more precise identification of optimal learning pathways for individuals and groups. Furthermore, we foresee the potential for these systems to not only adapt to individual learners but also to predict and prepare for future skill requirements based on emerging trends and technologies. This proactive approach to training could revolutionize workforce development and national security preparedness. However, as we move towards this AI-enabled future of training and education, it is crucial to maintain a human-centric approach. The ethical implications of AI in education, including issues of privacy, bias, and the changing role of human instructors, will need careful consideration and ongoing research.

REFERENCES

Alayrac, J. B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., ... & Simonyan, K. (2022). Flamingo: a Visual Language Model for Few-Shot Learning. arXiv preprint arXiv:2204.14198.

Anthropic. (2023). Anthropic's Constitutional AI: Claude. Anthropic Blog. https://www.anthropic.com/index/introducing-claude

Aptima. (2023). Instructional system design (ISD) analysis using AI large language models. Technical report. https://www.aptima.com/wp-content/uploads/2024/02/NAUTICAL-White-Paper.pdf

Beal, J., & Jain, S. (2021). Digital Twins for Personalized Training and Education in the Department of Defense. Proceedings of the Interservice/Industry Training, Simulation and Education Conference (I/ITSEC) 2021, Paper No. 21248. National Training and Simulation Association (NTSA).

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. arXiv preprint arXiv:2005.14165.

Born, W., Bruni, S., Detzler, L., & Yerdon, V. (2023). Applying the sidekick principles to the prototyping of human-machine teaming in collection planning. Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 67(1), 537-541. https://doi.org/10.1177/21695067231193688

Chaparro-Osman, M., Rebensky, S., Reinert, A., Yerdon, V., Jenkins, C., Logue, J., Jusko, C., Gangberg, G. (2023). Wires crossed in a digital world: How to prevent misalignments in human and AI decision making. In Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC) Orlando, FL.

Cottam, J.A., Glenski, M.F., Shaw, Z.H., Rabello, R.S., Golding, A.J., Volkova, S., & Arendt, D.L. (2021). Graph comparison for causal discovery. Visualization in Data Science 2021.

Department of Defense. (2020). DoD Instruction 1322.35: Military Training. https://www.esd.whs.mil/Portals/54/Documents/DD/issuances/dodi/132235p.pdf

Ezer, N., Bruni, S., Cai, Y., Hepenstal, S. J., Miller, C. A., & Schmorrow, D. D. (2019, November). Trust engineering for human-AI teams. In Proceedings of the human factors and ergonomics society annual meeting (Vol. 63, No. 1, pp. 322-326). Sage CA: Los Angeles, CA: SAGE Publications.

Fletcher, J. D. (2009). Education and training technology in the military. Science, 323(5910), 72-75.

Fouse, A., Levchuk, G., Schurr, N., McCormack, R., Pattipati, K., & Serfaty, D. (2019). Aligning Teams to the Future: Adapting Human-Machine Teams via Free Energy. In International Conference on Intelligent Human Systems Integration (pp. 471-477).

Fouse, A., Weiss, C., Mullins, R., Hanna, C., Nargi, B., & Keefe, D. F. (2018). Multimodal Interactions In Multi-Display Semi-Immersive Environments. In 2018 IEEE Conference on Cognitive and Computational Aspects of Situation Management (CogSIMA) (pp. 36-41). IEEE.

Gao, J., Galley, M., & Li, L. (2018). Neural approaches to conversational AI. Foundations and Trends in Information Retrieval, 13(2-3), 127-298.

Glenski, M., & Volkova, S. (2021). Identifying Causal Influences on Publication Trends and Behavior: A Case Study of the Computational Linguistics Community. In Proceedings of the First Workshop on Causal Inference and NLP (pp. 83-94).

Greff, K., Srivastava, R. K., Koutník, J., Steunebrink, B. R., & Schmidhuber, J. (2016). LSTM: A search space odyssey. IEEE transactions on neural networks and learning systems, 28(10), 2222-2232.

Guo, G., Glenski, M.F., Shaw, Z.H., Saldanha, E.G., Endert, A., Volkova, S., & Arendt, D.L. (2021). VAINE: Visualization and AI for natural experiments. IEEE VIS 2021.

Horawalavithana, S., Ayton, E., Sharma, S., Howland, S., Subramanian, M., Vasquez, S., ... & Volkova, S. (2022). Foundation Models of Scientific Knowledge for Chemistry: Opportunities, Challenges and Lessons Learned. Challenges & Perspectives in Creating Large Language Models, 160.

Horawalavithana, S., Ayton, E., Usenko, A., Cosbey, R., & Volkova, S. (2023). Anticipating Technical Expertise and Capability Evolution in Research Communities using Dynamic Graph Transformers. arXiv preprint arXiv:2307.09665.

Joint Chiefs of Staff. (2018). Joint Publication 3-0: Joint Operations. https://www.jcs.mil/Portals/36/Documents/Doctrine/pubs/jp3_0ch1.pdf

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. Science, 349(6245), 255-260.

Karakra, A., Fontanili, F., Lamine, E., & Lamothe, J. (2019). HospiT'Win: A predictive simulation-based digital twin for patients pathways in hospital. IEEE Access, 7, 37377-37390.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in Neural Information Processing Systems, 33, 9459-9474.

Lu, Y., Xiao, Y., & Sears, A. (2020). The impact of personalization on the effectiveness of chatbots: An exploratory study. International Journal of Human-Computer Studies, 138, 102402.

Madni, A. M., Madni, C. C., & Lucero, S. D. (2019). Leveraging digital twin technology in model-based systems engineering. Systems, 7(1), 7.

Mazhari, S., Garg, H., & Banerjee, P. (2021). Digital Twin-Driven Approach for Personalized Training and Skill Assessment. In 2021 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR) (pp. 302-305).

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. ACM Computing Surveys (CSUR), 54(6), 1-35.

NATO Science and Technology Organization. (2021). Science & Technology Trends 2020-2040: Exploring the S&T Edge. https://www.nato.int/nato_static_fl2014/assets/pdf/2021/3/pdf/210303-ST-Tech-Trends-Report-2020-2040.pdf

Nielson, P. E., & Kratiak, C. (2021). The future of military training: Blending live and virtual in a synthetic training environment. Military Medicine, 186(Supplement_1), 38-45.

Pearl, J. (2009). Causality: Models, Reasoning, and Inference (2nd ed.). Cambridge University Press.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., & Chen, M. (2022). Hierarchical Text-Conditional Image Generation with CLIP Latents. arXiv preprint arXiv:2204.06125.

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., ... & Norouzi, M. (2022). Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. arXiv preprint arXiv:2205.11487.

Saldanha, … S. Volkova. (2020). Evaluation of Algorithm Selection and Ensemble Methods for Causal Discovery. In Causal Discovery & Causality-Inspired Machine Learning Workshop at Neural Information Processing Systems.

Sammut, C. (2021). AI-Driven Digital Twins for Adaptive Learning. In M. Virvou, E. Alepis, G. A. Tsihrintzis, & L. C. Jain (Eds.), Machine Learning Paradigms: Advances in Deep Learning-based Technological Applications (pp. 133-151). Springer International Publishing.

Shresha, D. A. P., Maharjan, S., & Volkova, S. (2019). Forecasting social interactions from dynamic graphs: A case study of twitter, github, and youtube. In Proceedings of the 15th International Workshop on Mining and Learning with Graphs (MLG).

Shrestha, P., Maharjan, S., Arendt, D., & Volkova, S. (2019). Learning from dynamic user interaction graphs to forecast diverse social behavior. In Proceedings of the 28th ACM international conference on information and knowledge management (pp. 2033-2042).

Sikos, L. F., & Philp, D. (2020). Countering adversarial attacks in military AI applications. In 2020 IEEE 10th International Conference on Intelligent Systems (IS) (pp. 305-309). IEEE.

Sottilare, R. A., Burke, C. S., Salas, E., Sinatra, A. M., Johnston, J. H., & Gilbert, S. B. (2018). Designing adaptive instruction for teams: A meta-analysis. International Journal of Artificial Intelligence in Education, 28(2), 225-264.

Sottilare, R. A., Graesser, A. C., Hu, X., & Holden, H. K. (Eds.). (2018). Design recommendations for intelligent tutoring systems: Volume 6-team tutoring. US Army Research Laboratory.

Spirtes, P., Glymour, C. N., Scheines, R., & Heckerman, D. (2000). Causation, prediction, and search. MIT press.

Srivastava, S., Trehan, R., & Singh, S. (2019). A review on applications of artificial intelligence in e-commerce. International Journal of Business Intelligence and Data Mining, 15(3), 252-271.

Umbrello, S., & Van de Poel, I. (2021). Mapping value sensitive design onto AI for social good principles. AI and Ethics, 1(3), 283-296.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.

Volkova, S., Ayton, E., Porterfield, K., & Corley, C. D. (2017). Forecasting influenza-like illness dynamics for military populations using neural networks and social media. PloS one, 12(12), e0188941.

Volkova, S., et al. (2023). Explaining and predicting human behavior and social dynamics in simulated virtual worlds: reproducibility, generalizability, and robustness of causal discovery methods. Computational and Mathematical Organization Theory, 29(1).

Wager, S., & Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. Journal of the American Statistical Association, 113(523), 1228-1242.

Wooldridge, M. (2009). An introduction to multiagent systems. John Wiley & Sons.

Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., & Philip, S. Y. (2020). A comprehensive survey on graph neural networks. IEEE transactions on neural networks and learning systems, 32(1), 4-24.

Xin, D., Ma, L., Liu, J., Macke, S., Song, S., & Parameswaran, A. (2018). Accelerating human-in-the-loop machine learning: Challenges and opportunities. In Proceedings of the Second Workshop on Data Management for End-To-End Machine Learning (pp. 1-4). https://doi.org/10.1145/3209889.3209894

Zaharia, M., Khattab, O., Chen, L., Davis, J. Q., Miller, H., Potts, C., ... & Ghodsi, A. (2024). Compound AI Systems for Enhanced Training and Learning. arXiv preprint arXiv:2402.12345.

Zhou, J., Cui, G., Hu, S., Zhang, Z., Yang, C., Liu, Z., ... & Sun, M. (2020). Graph neural networks: A review of methods and applications. AI open, 1, 57-81.