Feature Article

Autonomous Intelligent Agents for Team Training

Making the Case for Synthetic Teammates

Christopher Myers
Air Force Research Laboratory

Jerry Ball
Oak Ridge Institute for Science and Education

Nancy Cooke
Arizona State University

Mary Freiman
Aptima

Michelle Caisse
L3 Technologies

Stuart Rodgers
TiER1

Mustafa Demir
Arizona State University

Nathan McNeese
Clemson University

***bstract*—The rise in autonomous system research and development combined with the maturation of computational cognitive architectures holds the promise of high-cognitive-fidelity agents capable of operating as team members for training. We report an ACT-R model capable of operating as a team member within a remotely piloted aerial system and provide results from a first-of-its-kind controlled randomized empirical evaluation in which teams that worked with an AST were compared against all-human teams. Our results demonstrate that ASTs can be incorporated into human teams providing training opportunities when teammates are unavailable. We conclude with issues faced in developing ASTs and lessons learned for future and current developers.**

Training is an integral part of the human ability to systematically progress socially and technologically. Training teams requires not only instructing individuals on their respective task(s), but also in when/how to interact with teammates. Currently, team training requires all members to be available, complicating the scheduling of training exercises. In large teams, having all team members present for training may be intractable, at which point *confederates* are brought in to *simulate* missing team members. This is costly as hiring confederates adds to the expense of training and does not address scheduling complexities. One solution is to develop intelligent agents capable of independently determining their own goals and strategies while acting with others to accomplish a

Digital Object Identifier 10.1109/MIS.2018.2886670
Date of publication 13 December 2018; date of current version 3 May 2019.

common team goal.[1] We refer to any intelligent agent that is meant to act as a team member as *autonomous synthetic teammates* (ASTs). ASTs could closely approximate human cognitive processes and behavior (e.g., computational cognitive models) or apply abstractions of human processes (e.g., deep reinforcement learning), or provide optimal behavior that results in less "human-like" behavior (e.g., control theoretic approaches).

ASTs used for training should closely approximate human-level cognitive processing–*high-cognitive-fidelity*. This premise is based on the assumption that training humans with ASTs capable of providing perfect information nearly instantaneously will produce lower levels of operational effectiveness because trainees will not be prepared to adapt to teammate mistakes or longer task execution durations that are associated with human performance. Consequently, we postulate that the use of a high-cognitive-fidelity AST is desirable, perhaps required, in training situations.

Cognitive models can come in many forms, with the most common being systems of systems like cognitive architectures (e.g., ACT-R).[2] Cognitive architecture is intended to account for invariant aspects of human cognition and provide a structure for integrating cumulative progress. Cognitive architecture is typically instantiated as software and serves as a foundation for the development of computational cognitive process models. Complex models, such as ASTs, depend on the integrated operation of many cognitive processes, and quickly become large-scale systems-of-systems models. Such models contain hypotheses about underlying cognitive mechanisms at the architectural level, such as times associated with memory retrievals, as well as hypotheses about the knowledge and strategic approaches that are brought to bear on a task.

We sought to objectively determine if teams with an AST approximate performance of all-human teams. Larger-scale models are significantly more difficult to evaluate as they may span times associated with elementary behaviors ($\approx$100 ms) to phenomena occurring over hours (e.g., learning), they tend to be nondeterministic, and the environments they tend to operate within provide opportunity for a variety of task strategies. Thus, ASTs are often evaluated via their ability to complete a desired task and/or through anecdotes from a subject matter expert, their developers, or their users during *in situ* demonstrations.[3] However, this does not suffice for objectively determining the differences between teams with and without an AST. Here, we report a first-of-its-kind evaluation of an AST capable of participating in team training scenarios as a teammate using a randomized control experiment within a remotely piloted aerial system (RPAS).

REMOTELY PILOTED AERIAL SYSTEM—SYNTHETIC TASK ENVIRONMENT

The RPAS—synthetic task environment (RPAS-STE) provides a test-bed for the study of team cognition for a three-person heterogeneous team.[4] In this STE, three participants coordinate to "fly" a simulated RPAS to take reconnaissance photographs. The RPAS-STE task was modeled after team task components from the United States Air Force Predator ground control station.[4] Three subjects are assigned to the role of pilot, photographer, or navigator. Individuals are first trained on the tasks specific to their roles and then come together to work as a team to complete five 40-min reconnaissance missions to photograph stationary ground targets. The task required teammates to communicate information necessary to successfully achieve the objective.

Each participant is seated in front of two computer monitors that display unique role information and common vehicle information (heading, speed, altitude). Team member interaction occurs through text-based communications similar to instant messaging and email, enabling the recording of sender/receiver identities and timing. Team and individual measures have been designed and validated and are embedded in the task software.[5] To objectively determine the team performance, a composite outcome score is computed for teams at the end of each 40-min mission based on the number of targets successfully photographed and the duration of warnings and alarms incurred. Data have been collected from eight different experiments in the RPAS-STE leading to the development of a theory of interactive team cognition.[5] Consequently, the RPAS-STE provides a well-understood task for developing and objectively evaluating ASTs.

AUTONOMOUS SYNTHETIC TEAMMATE

The AST was developed using the atomic components of thought—rational (ACT-R) computational cognitive architecture.[2] ACT-R was selected because it is a high-fidelity account of human cognitive capacities that can account for a broad coverage of human cognitive phenomena and the existence of a large scientific community continually contributing models and functionality (see http://act-r.psy.cmu.edu/). We first briefly cover the ACT-R cognitive architecture followed by the AST s core cognitive components.

Atomic Components of Thought—Rational

ACT-R is a hybrid cognitive architecture that includes continuous processes that operate over symbolic knowledge. Symbolic knowledge is discrete and divided into procedural knowledge that is implemented as IF–THEN rules (i.e., production rules) and declarative knowledge that represent retrievable facts (i.e., chunks). There are several cognitive capacities that are represented within ACT-R as buffers that can hold a single chunk at any time (e.g., declarative memory (DM), visual perception, control state). Productions that match buffer contents are selected to either affect the environment or ACT-R s buffer contents. Productions are selected and executed at a default cycle time of 50 ms. If more than one match the buffer contents, then the one with the greatest utility, *U*, is selected for execution

$$U_i(n) = U_i(n - \quad) + \quad R_i(n) - U_i(n - \quad)] \qquad (\ )$$

where the production utility U_i is the summation of the utility of a production   after its $n-1$th application, $U_i(n - \quad)$, and the reward of the production receives for its nth application $R_i(n)$. The learning rate   magnifies the reward receipt to either speed or slow learning.[2]

The ACT-R DM system is a bipartite architecture, where information is either in focus (in a buffer) or in long-term memory. To retrieve a memory, a request is made from a production. All declarative memories that match the request are considered for retrieval. The memory with the highest *activation* is selected

$$A_i = B_i + \sum_k \sum_j W_{kj} S_{ji} + \qquad (2)$$

where A_i is the activation for a chunk , B_i represents the base-level activation, reflecting the recency and frequency of practice of chunk . Activation is spread to chunk   from all chunks *j* in slots of chunks in buffer *k*, across all buffers, where W_{kj} is the amount of activation from source *j* in buffer *k*, and S_{ji} is the strength of association from source *j* to chunk . There is also noise added to the activation value . Thus, a chunk s activation is a reflection of its prior use B_i, activation spread from chunks in slots of chunks in buffers ${}_{k} \quad {}_{j} W_{kj} S_{ji}$, and noise . The chunk that has the highest activation that matched the retrieval request is returned if it is above a threshold. If the chunk is not above the threshold, then a retrieval failure occurs and the model has effectively forgotten the information. Using ACT-R to develop the AST situates our development focus on a high-cognitive-fidelity AST.

AST in ACT-R

The AST is one of, if not *the*, largest high-fidelity cognitive models ever built, containing over 2000 procedural memories and more than 57,000 declarative memories. Further, it operates over a timescale atypical for computational cognitive process models, performing five 40-min missions. (Milestones in the development of the AST have been previously documented in *Computational Mathematical Organization Theory*. For significantly greater detail, see the work of Ball *et al.*[6] and Rodgers *et al.*[7] The details from each of these manuscripts cannot be reproduced here for copyright and length restrictions.)

The AST was developed to pilot the RPAS and is divided into five components: the *natural language analysis* component, the *situation representation* component, the *agent–environment interaction* component, and the *dialog management* and *language generation* components. The language analysis component interacts with a situation representation component that contains spatial-imaginal/propositional representations of the current state of affairs encoded from reading text communications and interacting with the task environment. The situation representation component is intended to be a computational implementation of a situation model as described in the work of Zwaan and Radvansky[8] combined with Endsley s theory of situation awareness.[9,10] The dialog management and language generation components interact with the situation representation to determine when to say what and to whom. The agent–environment interaction component implements the "observable behaviors" of the system, controlling shifts of visual attention, and motor actions needed to communicate with teammates and to pilot the RPAS. Input to the AST is mediated by ACT-R s perceptual module and motor actions are mediated by ACT-R s motor module. In the following section, we provide more details for each of the five components of the AST.

Language Analysis Component The language analysis component is intended to be a domain general system capable of handling a broad range of English constructions. It is a construction-driven processing system based on a linguistic theory of the grammatical encoding of referential and relational meaning,[11] which is aligned with basic principles of cognitive and construction grammar (cf, Langacker[12,13]). Lexical items in the linguistic input activate constructions through ACT-R s DM system [see (2)] that drives further processing. The language analysis component is the largest component in the AST, containing 1567 production rules and 57,706 DM chunks.

The component adheres to two well-established cognitive constraints on language processing—incremental and interactive processing. The component processes the input incrementally (one word at time), constructing a linguistic representation of the input based on the current word, constructions activated by the word, and the prior context. If necessary, the current input is accommodated by adjusting the current representation or coercing the current input into that representation without backtracking or look ahead. The mechanism of context accommodation is a part and parcel of the basic left-to- right incremental processing mechanism.

The language analysis component is highly context sensitive and makes use of all information made available during text processing—lexical, syntactic, semantic, and pragmatic—in interactively deciding how to process a given input at each choice point. There is no independent syntactic component or syntactic processor, although grammatical information is very important for determining meaning. Contextual information is probabilistically summed via ACT-R s DM parallel spreading activation mechanism [see (2)] to yield the best alternative given the current input and context. The selected alternative is assumed to be correct and the processor proceeds deterministically and serially forward. The context-sensitive, probabilistic, parallel, spreading activation mechanism, combined with a mechanism of context accommodation makes possible a deterministic serial language processing system that builds a single representation. However, the system as a whole is pseudodeterministic in that the parallel integration of information at each choice point and the context accommodation mechanism are not characteristic of deterministic processing.

Situation Representation Component The situation representation is the component that functions as the primary interface between the other AST components and provides the primary meaning representation and inference capability of the overall system. A situation model is a mental representation of the propositional content of a text—including the addition of propositions corresponding to inferences that are derived from the text. The term situation model implies that this propositional representation is a model of the situation described within the text. For example, given the text "the next waypoint is LVN," a propositional representation like **hasAt tribute**(*RPAS*$_{01}$, *waypoint*$_{05}$) and **hasName**(*waypoint*$_{05}$, LVN) (where "the next" is resolved to refer to a specific order and "waypoint is LVN" is resolved to refer to a waypoint with a name of LVN) might be generated based on the grammatical pattern produced by language analysis.

The AST s situation representation reflects the dynamically integrated input from not only linguistic sources and discourse context, but also task processes and the model s knowledge.

Thus, the situation model represents propositional information both from reading text messages and from interacting with the task environment, providing an integration of linguistic situation models with the psychological construct of *situation awareness.*[9]

The situation representation is operationally defined as a set of objects, actions, events, and relationships associated with a task that is sufficient for reasoning about the agent s desired set of actions within the task. The situation representation is different from the agent s world knowledge but is related to, and affected by, world knowledge. World knowledge is provided to the AST *a priori* as a set of declarative facts, whereas the situation representation is a set of ACT-R buffers and declarative facts produced during operation from aspects of world knowledge, new task information, and text-based communications. Hence, the situation representation can be thought of as the synthetic teammate s mental model of the RPAS situation, including the RPAS, its flight parameters, its location relative to waypoints, the waypoints, etc. Just as the language analysis component incrementally builds a representation of a sentence, the situation model uses that evolving representation to build a situation instance.

Agent–Environment Interaction Component The agent–environment interaction component of the synthetic teammate was developed to directly interact with the point-and-click RPAS-STE interface. Flying to waypoints involves interacting with the RPAS-STE to queue the correct waypoint and enter the correct course. The pilot must also set the RPAS airspeed and altitude within restrictions provided by the photographer and navigator. The agent–environment interaction component interacts with the RPAS-STE using the same device as humans—it uses the mouse pointer to click on flight controls and uses the keyboard to send and receive messages to/from its teammates.

There is significant complexity in developing computational process models that interact with graphical user interfaces (GUIs) based on the need to model mouse pointer movement, attention selection, saccade times, etc. The use of the ACT-R architecture made it tractable to develop a model capable of interacting with the RPAS-STE GUI, as it provided built-in functionality for modeling shifts of visual attention, mouse pointer movements, button clicks, etc.

Language Generation and Dialog Management Components The language generation and dialog management components were developed to capture the dynamic nature of human language production, following earlier approaches involving dynamic dialog constraints, accommodation, and adaptive content selection. The focus of the language generation component is on selecting from a set of possible utterances, akin to overgeneration-and-ranking approaches. The focus of the dialog management component is the management of communication obligations with messages abstracted as dialog acts.

The language generation component uses optimality theory[14] to select an optimal utterance, given a set of utterances and their constraints. Constraints are simple, violable, conflicting, and motivated by cross-linguistic evidence. Constraints are arranged in a strict dominance hierarchy; the optimal utterance is the one that least violates the hierarchy. Constraint ranking is expressed through ACT-R DM activation: the most important constraint is most highly activated. Activation spreads from constraints to utterances to determine the utterance retrieved from memory; the most important constraint has the greatest effect on the retrieval. Factors from the situation model component dynamically affect the constraint ranking, providing a principled variation in utterances over time. Language generation is based on retrieval of complete messages with one or two variabilized slots. These message templates are akin to constructions, but there is currently no capability to integrate multiple constructions together, as in the language comprehension component. Purely constraint-based approaches like optimality theory are good at selecting among competing alternatives, but require additional mechanisms to support productive generation of alternatives.

The dialog management component models the push and pull of information to and from the synthetic teammate. Messages are abstracted as speech acts based on the DAMSL annotation scheme developed by Core and Collen.[15] For example, the message "Are there any restrictions for LVN?" is classified as a forward-looking check question. In normal conversation, there is a sense of obligation to follow rules of communication. For example, if a question is asked, then that question should be answered, or at least addressed somehow. Obligation rules facilitate effective discourse by setting up expectations for future messages. Messages with little local context information can then be understood because of the context from previous expectations. In the current dialog management component, obligations are stored as declarative chunks in a specialized module with separate buffers for self and others. Dialog management productions create, release, and use obligations to fill in context. In addition to the obligation module, the dialog management component also uses a temporal module extension to ACT-R to avoid repeatedly asking for the same information.

Executive Control Each component of the AST makes use of the same DM and production system provided by ACT-R. Consequently, there is a central processing bottleneck, the production system, as each component needs it to process information. This is managed with a simple *executive control* mechanism that enables the different AST components to obtain processing control once a component has relinquished control. For example, if the language analysis component is reading and processing text, then the agent–environment interaction component cannot process any information and vice versa. When a component completes processing, then it *relinquishes control* of the system by returning it to a "completed" goal state. At this point, the other components that have productions that match the "completed" goal state are free to *obtain control* and retrieve a new goal from DM. If more than one production matches the "relinquished control" state, then the production with the highest utility is selected [see (1)]. Consequently, the AST autonomously determines its own goals to pursue during mission execution through the dynamics of the ACT-R production utility and DM retrieval equations (see (1) and (2), respectively).

The situation representation component was the only component that did *not* adhere to the simple executive control mechanism because it was developed to maintain information derived from processing within each of the other components. In an effort to have a more thorough, complete situation representation, rules from the situation component could execute at any time, independent of which component had control of the production system. This, of course, comes at the cost of inadvertently interrupting other components at the risk of failing to complete an intended goal.

EMPIRICALLY EVALUATING THE AST

To determine if the AST could provide enough task skill and communication capabilities to facilitate behavior at the team and individual levels of analysis, we manipulated team composition using three subjects conditions: synthetic, experimenter, and control. The control condition was a task-naive all-human team. The experimenter condition had an expert human serve as the RPAS-STE pilot with task-naive human photographers and navigators. The synthetic condition had the AST performing as the pilot, also with task-naive human photographers and navigators.

In the experimenter condition, the expert pilot focused on effective coordination of information within the team. The role and instruction given to the pilot in this condition were the same as the other two conditions, the only difference being they were experienced at coordinating task-specific information within and among the team. Specifically, the pilot in this condition would push and pull information among the team members if information was not given after a set amount of time, or if it was not forthcoming. To ensure that coordination occurred in a structured and routine manner across all teams, the pilot used a coordination script. This script consisted of IF–THEN statements dependent on the task itself. For example: *If* the photographer does not request a certain airspeed or altitude of a reconnaissance target within one minute, *then* the pilot asks if the current speed and altitude are correct. The increased reliability of pushing and pulling information throughout the team in this condition is hypothesized to increase team performance. In addition, the experimenter condition provided a high-level benchmark for how an extremely high performing and *expert* AST should perform for training advanced teams. However, the AST pilot performance differed from a novice pilot provides an opportunity to better understand the weaknesses of the AST for focused improvement.

Individuals were randomly assigned to form teams of three and then randomly assigned to each condition. Each team completed five unique missions, with the last mission being one of high cognitive workload (many more targets than missions 1–4). There were 10 teams per team composition condition. To objectively determine the effects of the AST, performance was compared between conditions across individual and team reconnaissance tasks. We first present results from team.

Team Level Performance Analyses

The team performance score is a sum of penalties accrued during a mission subtracted from 1000 points. Each of these penalties was weighted to correspond with task importance. To determine how teams differed in their performance based on our three conditions and across missions, we performed a 3 (team composition) × 5 (missions) repeated measures multivariate ANOVA on the team performance score. Mauchly s test indicated that the assumption of sphericity was violated for mission, $\chi^2(9) = 22.110$, $p < 0.05$, $= 0.762$, thus the Greenhouse–Geisser correction for degrees of freedom for the mission effects.

Importantly, the mission × team composition interaction was not significant $F(6.092, 82.246) = 1.884$, $p = 0.092$, demonstrating team conditions did not significantly differ with increasing task experience. The team composition main effect was statistically significant, $F(2,27) = 11.496$, $p < 0.001$, $\eta^2 = 0.460$, *demonstrating a performance difference across the different t*eams (i.e., synthetic, experimenter, and control). There was also a significant main effect of mission, $F\,(3.046, 82.246) = 2.991$, $p < 0.05$, $^2 = 0.100$, demonstrating a performance decrease from Mission 4 to the high-load Mission 5.

To determine which team conditions differed to drive the main effect, we planned comparisons between each of the team composition conditions. The results of the pairwise-dependent *t*-test indicated that teams in the synthetic and control conditions were not significantly different ($M_{\text{Synthetic}} = 821.00$, $M_{\text{Control}} = 812.167$, $p = 0.286$), yet teams in the synthetic condition did have significantly lower scores than teams from the experimenter condition, ($M_{\text{Exp}} = 849.40$, $p < 0.05$). Further, and not surprising, teams in the experimenter condition had significantly higher scores than control teams. To summarize, synthetic and control team composition conditions performed the same across missions. However, the experimenter condition outperformed the other two.

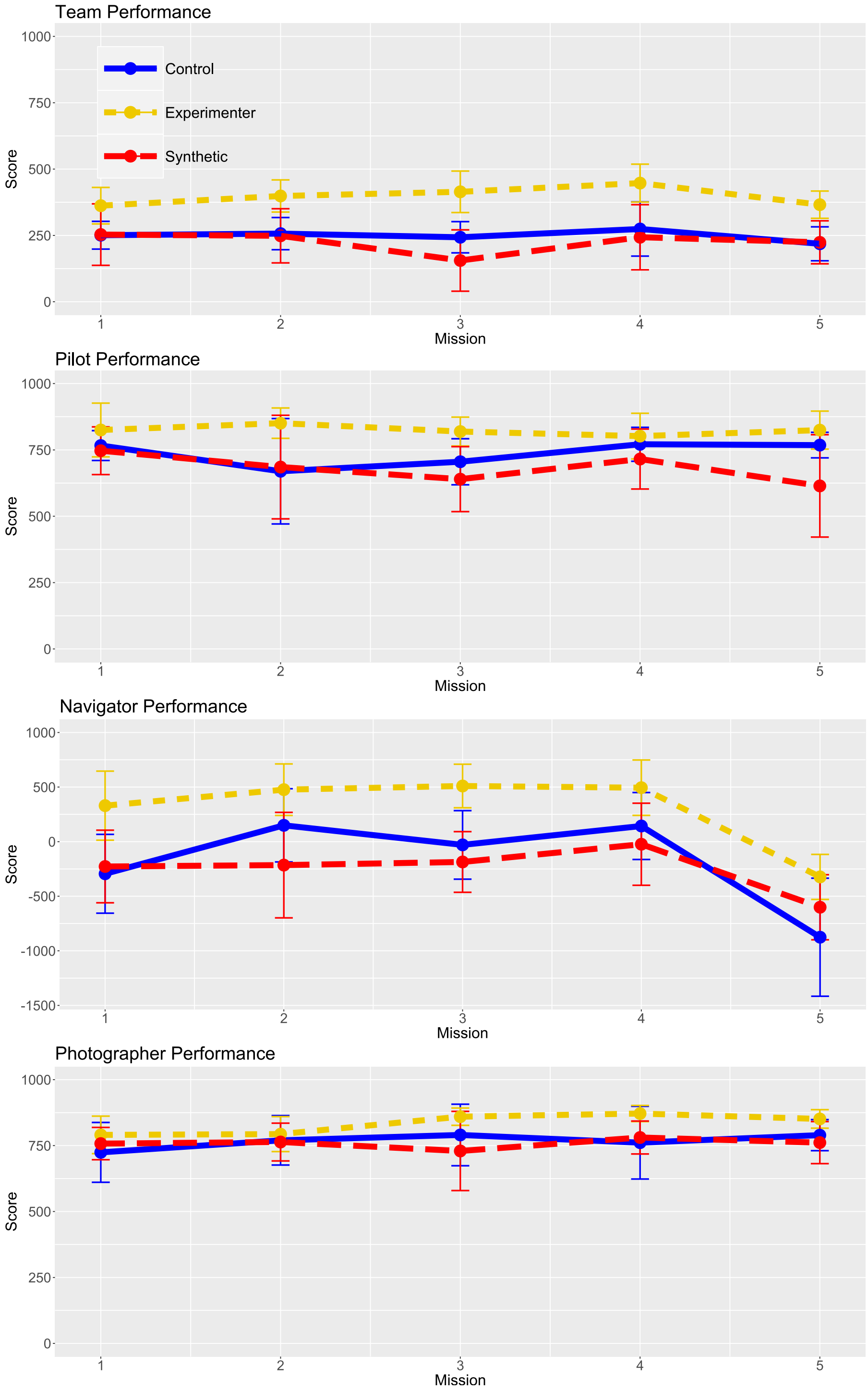
Task Level Performance Analyses

In the following sections, we analyze pilot, navigator, and photographer performance. Similar to the team performance scores, the individual task performance scores were computed by subtracting the sum of multiple penalties from 1000 points. We begin with the pilot, followed by the navigator, and complete the analyses with the photographers performance differences between team composition conditions and missions.

Pilot Performance The pilot performance score is composed of the sum of penalties accrued during a mission subtracted from 1000 points. We can directly compare how well the AST performed its piloting task relative to its human counterparts using the performance score. We hypothesized that control and experimenter teams would differ in their performance given that the experimenter had greater experience with piloting the RPAS, and that experimenter pilot performance would likely differ from control and synthetic conditions.

A 3 (team composition) × 5 (missions) repeated measures multivariate ANOVA on the pilot performance score was conducted to determine if there were significant differences across the different penalties and conditions. Mauchly s test indicates that the assumption of sphericity was violated for mission, $\chi^2(9) = 37.8$, $p < 0.001$, $= 0.56$. The Greenhouse–Geisser correction for degrees of freedom was thus used for the mission-based effects.

The results indicate that the condition main effect, $F(2, 27) = 4.88$, $p < 0.05$, $^2 = 0.27$, was significant, while the mission main effect, $F(2.24, 60.4) = 0.90$, $p = 0.42$, and the condition × mission



Team Performance

Pilot Performance

Navigator Performance

Photographer Performance

Figure 1. Pilot, navigator, photographer, and team performance scores for all human teams and teams with the synthetic teammate. Mission 5 is high cognitive load. Error bars are 95 confidence intervals. These results indicate that teams with the AST (top), and navigators and photographers working with the AST perform no differently than novice human teams, but underperforms relative to teams with an expert pilot (experimenter condition: bottom two plots). Further, the AST performs the piloting task as well as novice humans, but does not perform as well as an expert pilot. The results provide an objectively evaluated proof of concept of using ASTs within teams and maintaining its teammates performance.

interaction, $F(8,\ 108) = 0.99$, $p = 0.44$, were not significant. The pairwise test results indicate that the experimenter teams ($M_{\mathrm{Exp}} = 824$, $p < 0.05$) performed better than the synthetic ($M_{\mathrm{Synthetic}} = 680$, $p < 0.05$) and slightly better than the control teams, ($M_{\mathrm{Control}} = 736$, $p = 0.069$). The synthetic teams performance was not significantly different from the control teams performance, ($p = 0.24$). The teams with an expert pilot performed better than control and synthetic pilots, and there was no statistical difference between the control pilots and synthetic pilots (see Figure 1).

Navigator Performance The navigator performance score is also composed of the sum of penalties accrued during a mission subtracted from 1000 points. Like the pilot score, we can directly compare how well the navigators that interacted with the synthetic teammate performed their task relative to navigators that interacted with human pilots.

We conducted a 3 (team composition) × 5 (missions) repeated measures multivariate ANOVA on the navigator performance score Mauchly s test of sphericity indicated that the assumption was not violated for mission, $\chi^2(9) = 15.52$, $p = 0.078$, $\ = 0.791$. The mission × team composition interaction was not significant, $F(8,\ 108) = 1.433$, $p = 0.191$. The main effect of team composition was statistically significant, $F(2,\ 27) = 7.441$, $p < 0.05$, $\ ^2 = 0.355$, as was the mission main effect, $F(4,\ 108) = 25.614$, $p < 0.001$, $\ ^2 = 0.487$.

We investigated where the differences were between the three conditions. The results of the planned pairwise-dependent t-test indicated that navigators in the synthetic condition had no difference in performance scores than navigators in the control condition ($M_{\mathrm{Synthetic}} = 749.88$, $M_{\mathrm{Control}} = 763.74$, $p = 0.658$). However, the experimental condition had a greater score ($M_{\mathrm{Exp}} = 859.49$) relative to the navigators in the synthetic ($p < 0.001$) and control conditions ($p < 0.05$).

Photographer Performance Similar to the navigator performance, we can directly compare how well the photographer performed when the AST was piloting relative to when it was not piloting the RPAS. The photographers performance score is the summation of penalties subtracted from 1000 during a mission. We conducted a 3 (team composition) × 5 (missions) repeated measures multivariate ANOVA on the photographer performance score. Mauchly s test of sphericity indicated that the assumption was not violated for mission, $\chi^2(9) = 14.819$, $p = 0.097$, $\ = 0.801$.

The team composition main effect was not statistically significant, $F(2,\ 27) = 1.892$, $p = 0.170$ ($M_{\mathrm{Synthetic}} = 951.71$, $M_{\mathrm{Control}} = 953.42$, $M_{\mathrm{Exp}} = 966.66$). The main effect of mission was not significant, $F(4,\ 108) = 1.591$, $p = 0.182$, nor was the mission × team composition interaction, $F(8, 108) = 0.970$, $p = 0.463$.

Task Level Performance Summary There are two important points worth summarizing from the task level performance analyses: 1) the AST performed the piloting task at a level that was statistically indistinguishable from human RPAS pilots in the control condition; and 2) the navigators and photographers that interacted with the AST pilot performed at a level that was statistically indistinguishable from those that interacted with a human pilot.

CONCLUSIONS AND DISCUSSION

Teams with the AST performed as well as the control teams, yet not as well as the experimenter teams. Importantly, teammates that had to complete missions with the synthetic teammate (i.e., navigators and photographers) performed their tasks at a level of performance statistically indistinguishable from navigators and photographers that worked with naive human pilots (i.e., the control condition). This is an important demonstration of a synthetic teammate supplanting a human operator and maintaining team and individual effectiveness. While this is a first and necessary step, we must next directly address the question of training by testing navigator and photographer in all-human teams after learning the task with an AST.

From the perspective of the AST as a computational cognitive model, it provides a fantastic zero-parameter fit—no parameters were manipulated and retested to improve either pilot or team performance scores after collecting team data. This is remarkable when you consider the AST is composed of over 2000 procedural and 57,000 declarative memories and operates over multiple 40-min missions.

The widespread development and distribution of ASTs is on the horizon, be they digital assistants, autonomous drivers, or military systems. However, achieving an expansion of research and development on ASTs, and thus greater deployment, hinges on overcoming four significant challenges identified during our AST research: 1) rapidly developing ASTs; 2) rapidly debugging and extending ASTs; 3) verifying and validating of ASTs; and 4) sustaining deployed ASTs. Each challenge follows from the requirements that ASTs operate within complex tasks and environments and naturally communicate with human teammates while achieving targeted levels of functionality.

It remains unclear if ASTs for training must be high cognitive fidelity. If the current AST did not closely approximate human behavior, such as times for clicking buttons, completing piloting subtasks, and piloting unit tasks, then could the AST s approximation to human behavior be reduced further yet maintain team and individual performance? This is an open and outstanding empirical question.

The ability to reduce cognitive fidelity while maintaining team and individual performance would reduce AST complexity and facilitate the ease and rate of AST development, thus increasing their prevalence in society (Challenge 1). An important milestone in AST development, verification and validation (Challenge 3) within a controlled empirical study, was reported here, and we believe this is the first reported in the scientific literature. The ability to easily debug, extend, and sustain deployed ASTs remain as challenges for future focus.

ACKNOWLEDGEMENT

The work of C. Myers was supported by the Office of Naval Research (N000140910201). The work of N. Cooke was supported by the Office of Naval Research (N000141110844). The authors would like to thank the Air Force Research Laboratory for their continuing support in AST research.

REFERENCES

1. W. Zachary, T. Santarelli, D. Lyons, M. Bergondy, and J. Johnston, “Using a community of intelligent synthetic entities to support operational team training,” in *Proc. 10th Conf. Comput. Generated Forces Behavioral Representations*, 2001, pp. 215–233.
2. J. R. Anderson, *How Can the Human Mind Exist in the Physical Universe?*, F. E. Ritter, Ed. London, U.K.: Oxford Univ. Press, 2007.
3. J. E. Laird and R. M. Jones, “Building advanced autonomous AI systems for large scale real time simulations,” in *Proc. Comput. Games Develop. Conf.*, Long Beach, CA, USA, 1998, pp. 365–378.
4. N. J. Cooke and S. M. Shope, “Designing synthetic task environments,” In S. G. Schiflett, L. R. Elliott, E. Salas, & M. D. Coovert, *Scaled Worlds: Development Validation and Application*, Surrey, England: Ashgate, pp. 263–278, 2004.
5. N. J. Cooke, J. C. Gorman, C. W. Myers, and J. L. Duran, “Interactive team cognition,” *Cogn. Sci.*, vol. 37, pp. 255–285, 2013.
6. J. Ball et al., “The synthetic teammate project,” *Comput. Math. Org. Theory*, vol. 16, no. 3, pp. 271–299, Aug. 2010.
7. S. Rodgers, C. Myers, J. Ball, and M. Freiman, “Toward a situation model in a cognitive architecture,” *Comput. Math. Org. Theory*, vol. 19, pp. 313–345, 2013.
8. R. Zwaan and G. Radvansky, “Situation models in language comprehension and memory,” *Psychol. Bull.*, vol. 123, no. 2, pp. 162–185, 1998.
9. M. Endsley, “Toward a theory of situation awareness in dynamic systems,” *Human Factors*, vol. 37, pp. 32–64, 1995.
10. M. R. Endsley, “Situation awareness misconceptions and misunderstandings,” *J. Cogn. Eng. Decis. Making*, vol. 9, no. 1, pp. 4–32, Feb. 2015.
11. J. Ball, “A bi-polar theory of nominal and clause structure and function,” *Annu. Rev. Cogn. Linguistics*, vol. 5, no. 1, pp. 27–54, 2007.
12. R. W. Langacker, *Foundations of Cognitive Grammar: Theoretical Prerequisites*, vol. 1. Stanford, CA, USA: Stanford Univ. Press, 1987.
13. R. W. Langacker, *Foundations of Cognitive Grammar: Descriptive Application*, vol. 2. Stanford, CA, USA: Stanford Univ. Press, 1991.
14. A. Prince and P. Smolensky, “Optimality theory: Constraint interaction in generative grammar,” *Stud. Second Lang. Acquisition*, vol. 28, no. 1, pp. 1–262, 1993.
15. M. G. Core and J. F. Allen, “Coding dialogs with the DAMSL annotation scheme,” in *Proc. AAAI Fall Symp. Communicative Action Humans Mach.*, 1997, pp. 28–35.

Christopher Myers seeks to understand how goal-directed cognition is shaped by the accommodation of basic interactive processes to the statistical structure of the environment. His research is diverse, spanning processes fundamental to perceptual encoding and motor actions to adaptive human–machine team decision-making in nonstationary environments to the development and evaluation of high-cognitive-fidelity synthetic teammates. He uses computational cognitive process and mathematical models to achieve his research and development goals, and is well versed in Bayes theorem, the Atomic Components of Thought–Rational (ACT-R), and soar modeling frameworks. His long-term research goal is to deploy computational cognitive process models as teammates that facilitate training and operations. He became a member of the Cognitive Science, Models, and Agents branch of the Warfighter Readiness Research Division in 2009, where he continues to perform research and lead the Cognitive Models Core Research Area. He received the Ph.D. degree in cognitive science from Rensselaer Polytechnic Institute in 2007, and was a postdoctoral scholar with Dr. N. Cooke from 2007 to 2009. He is the corresponding author. Contact him at christopher.myers.29@us.af.mil.

Jerry Ball currently has a fellowship at the Oak Ridge Institute for Science and Education under the Knowledge Preservation Program. Previously, he was a senior research psychologist in the Air Force Research Laboratory, 711th Human Performance Wing, Warfighter Readiness Research Division, Cognitive Science, Models, & Agents Branch until November 2015 when he retired. He was the technical lead on a project to develop a synthetic teammate capable of functioning as the pilot in a three-person simulation of a UAV reconnaissance mission. This capability was validated by the first of its kind empirical evaluation of a synthetic teammate interacting with human teammates and achieving a level of team performance that equaled or exceeded the team performance of all human teams. Contact him at jerryandauroraball@gmail.com.

Nancy Cooke is a professor of human systems engineering at Arizona State University and is the science director of the Cognitive Engineering Research Institute in Mesa, AZ, USA. She also directs ASU's Center for Human, AI, and Robot Teaming, and the Advanced Distributed Learning DOD Partnership Lab. She is the immediate past president of the Human Factors and Ergonomics Society and the recent past chair of the Board on Human Systems Integration at the National Academies of Science, Engineering, and Medicine. She received the Human Factors and Ergonomics Society's Arnold M. Small President's Distinguished Service Award in 2014. Her research interests include the study of individual and team cognition and its application to the development of cognitive and knowledge engineering methodologies, human–robot teaming, sensor operator threat detection cyber and intelligence analysis, remotely piloted aircraft systems, healthcare systems, and emergency response systems. She received the Ph.D. degree in cognitive psychology from New Mexico State University in 1987. Contact her at nancy.cooke@asu.edu.

Mary Freiman is a senior research scientist with Aptima, Inc. She has over a decade of experience building language-capable synthetic agents using computational cognitive architectures and programming tools. Her research has focused on the modeling of situation awareness, reading comprehension, decision-making tasks, and the evaluation of cognitive architectures in the context of simulation and agent building. She applies the insight gained from cognitive models to improve how human teams interact with autonomous synthetic teammates. She received the B.A. degree in linguistics and the M.S. degree in human language technology both from the University of Arizona. Contact her at mfreiman@aptima.com.

Michelle Caisse is a computational linguist with strong software development skills. Before coming to Link Simulation and Training and the Air Force Research Laboratory Warfighter contract, she was involved in developing speech technology software at Berkeley Speech Technologies and in software quality engineering at Sun Microsystems and other companies. At AFRL, she worked the eXpert Common Immersive Threat Environment (XCITE)and was the project lead for the XCITE Radio Frequency Threat Environment (RFTE) project, which incorporated multi-language, dynamic digital audio simulation of radio communications into XCITE and, following that, into the Network Integrated Constructive Environment (NICE). She then joined the Synthetic Teammate project working on the dialog and language generation components of the synthetic agent.

Stuart Rodgers has been participating in research projects across multiple Department of Defense Services laboratories, NASA, and DARPA since 2001. These projects focused on human performance

modeling and analysis. He has authored and presented award-winning scientific papers on heuristic search, computational cognitive process models, human performance modeling, and knowledge representation topics. Prior to TiER1, he served with the U.S. Air Force as a flight instructor and an experimental test pilot and held a number of leadership positions in Air Force research, development, and test offices. He is a senior member of the IEEE Computer Society. He received the B.S. degree in aeronautical engineering from the U.S. Air Force Academy, the M.S. degree in computer science from the University of West Florida, and a Management of Technology Certificate from the California Institute of Technology. He is a graduate of the Experimental Test Pilot Course at the U.S. Air Force's Test Pilot School. Contact him at s.rodgers@tier1performance.com.

Mustafa Demir is currently a postdoctoral research associate working in direct collaboration with Dr. N. Cooke at Ira A. Fulton Schools of Engineering, Arizona State University. His current research interests include human-autonomy teaming, team cognition, complex systems, and advanced statistical modeling. He received the Ph.D. degree in simulation, modeling, and applied cognitive science with a focus on team coordination dynamics in human-autonomy teaming from Arizona State University in Spring 2017. Contact him at mdemir@asu.edu.

Nathan McNeese is an assistant professor and the director of the Team Research Analytics in Computational Environments (TRACE) Research Group within the division of Human-Centered Computing in the School of Computing, Clemson University. He also holds a secondary appointment in Clemson's Human Factors Institute, is a Faculty Scholar in Clemson's School of Health Research, and is a Watt Family Faculty Fellow. His research interests and expertise include human–machine teaming, the study of team cognition and technology, and the development/design of human-centered collaborative tools and systems. He received the Ph.D. degree in information sciences and technology with a focus on team decision making and cognition from The Pennsylvania State University in the fall of 2014. Contact him at mcneese@clemson.edu.