AFRL-RH-WP-TR-2019-0031

DYNAMICALLY MANAGING TASK ALLOCATION BETWEEN HUMANS AND MACHINES IN SURVEILLANCE OPERATIONS

Mary Frame
Alan Boydstun

Wright State Research Institute
4035 Colonel Glenn Hwy
Beavercreek, OH 45431

Jennifer Lopez

711th HPW Air Force Research Laboratory

August 2019
Interim Report

AIR FORCE RESEARCH LABORATORY
711TH HUMAN PERFORMANCE WING,
AIRMAN SYSTEMS DIRECTORATE,
WRIGHT-PATTERSON AIR FORCE BASE, OH 45433
AIR FORCE MATERIEL COMMAND
UNITED STATES AIR FORCE

NOTICE AND SIGNATURE PAGE

Using Government drawings, specifications, or other data included in this document for any purpose other than Government procurement does not in any way obligate the U.S. Government. The fact that the Government formulated or supplied the drawings, specifications, or other data does not license the holder or any other person or corporation; or convey any rights or permission to manufacture, use, or sell any patented invention that may relate to them.

This report was cleared for public release by the 88th Air Base Wing Public Affairs Office and is available to the general public, including foreign nationals. Copies may be obtained from the Defense Technical Information Center (DTIC) (http://www.dtic.mil).

AFRL-RH-WP-TR-2019-0031 HAS BEEN REVIEWED AND IS APPROVED FOR PUBLICATION IN ACCORDANCE WITH ASSIGNED DISTRIBUTION STATEMENT.

OCAMPO.SABRINA.M.1185673517 Digitally signed by OCAMPO.SABRINA.M.1185673517 Date: 2019.08.13 10:33:18 -04'00'

SABRINA OCAMPO, Work Unit Manager
Human Analyst Augmentation Branch
Airman Systems Directorate
711th Human Performance Wing
Air Force Research Laboratory

SIMPSON.RICHARD.D.1091443429 Digitally signed by SIMPSON.RICHARD.D.1091443429 Date: 2019.08.20 13:00:41 -04'00'

RICHARD D. SIMPSON, DR-IV, DAF
Human Centered ISR Division
Airman Systems Directorate
711th Human Performance Wing
Air Force Research Laboratory

This report is published in the interest of scientific and technical information exchange, and its publication does not constitute the Government's approval or disapproval of its ideas or findings.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
05 08 19	Interim Report	June 2017 – January 2019

4. TITLE AND SUBTITLE
Dynamically Managing Task Allocation between Humans and Machines in Surveillance Operations

5a. CONTRACT NUMBER
FA8650-12-D-6583 0002

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
00000F

6. AUTHOR(S)
*Mary Frame
**Jennifer Lopez
*Alan Boydstun

5d. PROJECT NUMBER
0000

5e. TASK NUMBER
00

5f. WORK UNIT NUMBER
H0L5

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
*Wright State Research Institute
4035 Colonel Glenn Hwy
Beavercreek, OH 45431

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
**Air Force Materiel Command
Air Force Research Laboratory
711th Human Performance Wing
Airman Systems Directorate
Human-Centered ISR Division
Human Analyst Augmentation Branch
Wright-Patterson AFB, OH 45433

10. SPONSORING/MONITORING AGENCY ACRONYM(S)
711 HPW/RHCM

11. SPONSORING/MONITORING AGENCY REPORT NUMBER(S)
AFRL-RH-WP-TR-2019-0031

12. DISTRIBUTION/AVAILABILITY STATEMENT
Distribution A: Approved for public release.

13. SUPPLEMENTARY NOTES
PA Clearance: MSC/PA-2018-0376; 88ABW-2018-6274, cleared on 4 January 2019

14. ABSTRACT

The purpose of this technical report is to articulate the construction of the Autonomous Manager (AM) as an integral component of distributing multiple tasks between humans and autonomous agents, particularly in Intelligence, Surveillance, and Reconnaissance (ISR), and provides information on the parameterization and validation of this tool via initial simulations prior to instantiation and empirical testing. These simulation studies were essential for calibrating the decision logic of the AM for scenarios where multiple simultaneous independent tasks must be maintained, for example, an analyst watching multiple unrelated Full-Motion Video (FMV) windows while also categorizing a series of Synthetic Aperture Radar (SAR) images. This provided a framework for understanding, as well as code that has been further tested in an empirical series of studies

15. SUBJECT TERMS
Human-Machine Systems; Intelligence Surveillance and Reconnaissance. Check DTIC Thesaurus

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT:	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON (Monitor)
a. REPORT	b. ABSTRACT	c. THIS PAGE			Sabrina OCampo
Unclassified	Unclassified	Unclassified	SAR	42	19b. TELEPHONE NUMBER *(Include Area Code)* N/A

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std. Z39-18

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1.0 SUMMARY

Increasingly sophisticated technology must be leveraged in surveillance environments to enable the eventual goal achievement of allowing analysts to increase throughput through management of multiple simultaneous feeds. Maintaining this increased tasking will likely introduce additional workload and fatigue. Fortunately, analysts can currently offload some of these tasks to automation, and will in the future be able to offload additional tasking to streamline the intelligence analysis process. Currently, various speech-to-text and text-to-speech programs can be used to convert spoken information into chat and automation can be used to copy text to multiple needed locations simultaneously. Automation has aided in the transmission of information between analysts and organizations. Tools are also being developed to augment the detection of important visual features within surveillance scenes. However, the degree of assistance autonomous systems can provide is still somewhat limited for cognitively complex tasks, but progress is being made incrementally toward viable assistive tools. Balancing analyst workload while maintaining multiple tasks will require intelligent and dynamic distribution of tasks between humans and autonomy.

To address this challenge of maintaining performance in a Human-Machine Team (HMT), we developed a program to dynamically distribute tasks between a human and automation. By re-conceptualizing the team dynamic within the surveillance working environment, we developed a supervisory role called the Autonomous Manager (AM). The AM dynamically reallocates tasks based on task performance and physiological indicators of a human analyst's workload. We tested the AM's decision logic across multiple scenarios using simulation, allowing us to examine the benefits and limitations of the AM more thoroughly than would be feasible with a series of empirical studies using human subjects. We tested the benefits of the AM based on: performance improvement across tasks, improvement with highly variable mean performance, and specific use-case scenarios. The simulated AM can be leveraged to answer a variety of real-world questions without the expense of physical implementation, prior to full-scale development or empirical testing.

2.0 THEORETICAL PERSPECTIVE ON HUMAN-MACHINE TEAMING

In most work environments, accuracy and efficiency are both crucial. To continue growing productivity and throughput, work organizations must increasingly rely on sophisticated automation to complete complex tasks. However, many cognitively taxing tasks cannot be performed by automation to a level of proficiency that matches an expert human at full arousal (i.e., an analyst who is not heavily fatigued or overworked). Integration of automation into real-world workplaces demands effective HMT to compensate for the relative strengths and weaknesses of each human and machine agent. Prior autonomy research has attempted to establish an appropriate balance of tasks from a range of theoretical perspectives. The perspective taken for development of our AM tool was from previous work in managing multiple tasks with humans and machines with twofold assumptions:

- Tasking can be distributed according to a spectrum, ranging from automation/machine having full control to human having full control. There are variable degrees of split tasking between humans and automation/machines that can include parsing complete or partial tasks between agents (Sheridan & Verplank, 1978; Parasuraman, Sheridan, & Wickens, 2000).
- Dynamic implementation of automation provides a greater benefit than simple substitution of a single or multiple humans. There are times where human control of tasks may be more appropriate and other times where automation control of tasking is more appropriate, based on features such as workload or fatigue (Parasuraman & Wickens, 2008).

2.1 Continuum of Task Control

Research over the past decades within autonomy has established many perspectives on how Human-Machine Teaming occurs based on how tasks are distributed, ranging from tasks that are entirely controlled by humans to tasks entirely controlled by automation (Parasuraman et al., 2000; Parasuraman & Wickens, 2008). Between these extremes are degrees of divided control over tasks. For example, supervisory control typically consists of an automated system with most of the direct control over specific task operations, but a human agent can interject and choose to veto the actions of the automation or take control manually (Kaber & Endsley, 2004). Likewise, a human may typically operate a task with automated intervention only in emergencies, as is the case of the Traffic Collision Avoidance System (TCAS) in modern commercial aircraft, which automatically adjusts elevation of two aircraft when they are about to collide, overriding pilot manual control until a collision is fully avoided (Feigh, Dorneich, & Hayes, 2012).

In one of the earliest models of a spectrum of task distribution, Sheridan and Verplank (1978) proposed a multi-level scale of HMT ranging from 1 (human makes decisions and controls tasking with no computer intervention) to 10 (automation makes decisions and controls tasking with no human intervention). Parasuraman et al. (2000) proposed a streamlined spectrum that

incorporates the functionality of the 10-level model, but simplifies it to 4 levels of automation based on cognitive functions: Sensory Processing, Perception/Working Memory, Decision Making, and Response Selection. These functions are based on human information processing and can provide an initial categorization for tasks: Information Acquisition, Information Analysis, Decision and Action Selection, and Action Implementation (Billings, 2018).

Patterson (2017) ascribed an additional cognitive explanation to these levels. Information acquisition and information analysis engages more intuitive processes while decision making and action implementation involve more analytic cognitive processes. Intuitive processes are typically rapid and automatic, while analytic processes are more deliberate, slow, and volitional (Evans, 2008). This distinction means that intervention by automation at different cognitive processing stages will likely have differentiable downstream effects on task performance. These differences may be particularly pronounced when certain tasks are differentially affected by time pressure and workload. For example, automation performing a more quickly-processed, intuitive (automatic) perceptual information gathering task may disrupt human performance on a later decision making task requiring integration of this perceptual information.

2.2 Adaptive Automation

Construction of effective automation depends not only on the calibration of human versus automation proportional control, but there is an additional benefit of adaptive automation that modifies functionality based on changing task environments (Parasuraman, Mouloua, & Molloy, 1996) and physiological indicators of human workload or fatigue (Byrne & Parasuraman, 1996). Furthermore, effective HMTs should involve some degree of monitoring or communication between agents to compensate for failures by either the human or automated agent (Parasuraman et al., 1996). Adapting the degree of involvement of automation can allow it to be used advantageously when human performance is inadequate, while reducing the “out-of-the-loop problem” which occurs when a human is completely disengaged from tasks other than to correct occurrences of catastrophic automation failure. Adaptive automation allows the human teammate to maintain Situation Awareness (SA) and flexibility to respond effectively to unexpected problems. Parasuraman et al. (1996) found that adaptive task allocation based on either a model or performance yielded significantly better detection of automation failures. Furthermore, in addition to making automation adaptive based on performance, automation has successfully been adjusted using psychophysiological data, such as electroencephalography (EEG), including the P300 event related potential (Prinzel III, Freeman, Scerbo, Mikulka, & Pope, 2003). The benefit of measures such as EEG, eye tracking, or physiological variables (e.g., heart rate, respiration rate), is that they can be obtained continuously over the course of a task, provide reaction in near real-time, and they can be observed even in the absence of an overt behavior or response. Even if a participant is unable to accurately assess their own workload state (either subjectively underestimating or overestimating workload), the physiological indicators are more likely to give an accurate assessment of cognitive state. Additionally, automation that is calibrated based on a human teammate’s physiological variation during an ongoing task has been shown to yield

consistently better human performance than when automation is implemented based on the identical pattern of variation when it is decoupled from psychophysiology (Prinzel, et al., 2003).

2.3 Practical Issues of Human-Machine Teaming in ISR

Prior research has formalized adaptive human-automation interactions within a single complex task environment. However, typical working environments require simultaneous management of multiple formal or informal tasks, rather than monotasking. Even within ISR, there is a push toward developing technologies to allow a single analyst to maintain superior performance while managing multiple simultaneous cognitively taxing efforts. If a human can manage multiple tasks adequately when working alone, it is unnecessary to develop expensive automation to replace them. However, physiological states such as overwork or fatigue can have a dramatic, typically negative, influence on human performance (Diekfuss, Ward, & Raisbeck, 2017; Hancock, Williams, & Manning, 1995), and this is when automation may be most effectively leveraged. Currently, the effectiveness of an ISR analyst managing multiple feeds without assistance is not quantified. Generally however, increased workload and performance decrements are observed when people must manage multiple tasks that require the same cognitive resources or modality (Stevens, Fisher, Morris, Myers, Spriggs, & Dukes, 2018). Watching any single screen would take away attentional and cognitive resources that could be allocated to watching the other screens. We can reasonably surmise that there could be a decrement in performance if there is no automated assistance when a human must devote a single cognitive modality to multiple tasks.

Over the past decade, there has been substantial progress in introducing automation into ISR workflows, such as Automatic Target Recognition (ATR) (Irvine & Nelson, 2009) for recognizing important target categories based on their features and systems to assist with reducing data and information complexity so the most pertinent information can be ascertained by analysts (Hershey, Wang, Graham, Davidson, Sica, & Dudash, 2012; Hershey & Wang, 2013). Although automation is being developed and pushed forward to manage complex tasks, it is not yet capable of consistently exceeding performance of a cognitively alert and expert human for certain actions, including making call-outs, noting higher-order patterns of behavior, and meaning-making. There is a tradeoff of performance between various subtasks within ISR. Therefore, automation should be implemented adaptively to appropriately allocate various subtasks in multitasking ISR environments to be most effective. The necessary level of control may change over time differentially between tasks, requiring real-time management of human and automated agents in task allocation. Knowing when to change tasks can be driven by measures of workload informed by ongoing collection of physiology, brain activity, and eye movement. Although variables including eye tracking metrics (e.g., blink rate, percentage of eyelid opening, saccade frequency) and physiology (e.g., heart rate, heart rate variability, EEG) can serve as workload indicators (Buettner, 2013; Gable, Kun, Walker, & Winton, 2015; Hoover, Singh, Fishel-Brown, & Muth, 2012; Luque-Casado, Perales, Cárdenas, & Sanabria, 2016; Palinko, Kun, Shyrokov, & Heeman, 2010), it is not possible for a human to integrate this

information and make appropriate decisions for task allocation in real-time. In intelligence and surveillance analysis however, this information is important and this managerial role could improve performance across multiple tasks.

2.4 Goal of Testing the Autonomous Manager

For this series of simulation studies, our primary goal was to test the benefits and limitations of the AM. These simulated efforts have been instrumental in the development of the AM prior to empirical testing. For this multi-task visual search simulation, the AM assessed four concurrent surveillance tasks (see Frame, Boydstun, Maresca, & Lopez, 2019 for an empirical implementation of the AM using four simultaneous static image search tasks). However, there are additional benefits of simulation beyond simple tool development. Simulations can perform the most difficult element of empirical research in an efficient manner and allows for a larger exploration of the sample space. Performing this series of studies as simulations allowed us to explore unlikely performance scenarios and test the robustness of our algorithm to dynamic and unpredictable tasking environments. Simulations also allow us to determine the upper bound of performance for potential HMTs using the AM versus baseline based on all-human task configurations, to determine where the greatest improvement can be made. There is also an inherent benefit to a simulation in practical application as a means of saving resources for proposed new tasking setups. This simulation allows for testing the marginal benefit of an AM, given a variety of proposed task arrangements. A human supervisor that can test and know where the greatest benefit can be achieved in advance can make wiser implementation choices for working environments and is reminiscent of other successful forms of adaptive automation informing a tasking environment (Miller & Parasuraman, 2007).

3.0 PRE-EXPERIMENTAL MATHEMATICAL VALIDATION

Prior to full development of a task allocation program, we needed to validate the assumption that an effective human machine team could consistently elicit superior performance to either agent (human or machine) performing multiple tasks in isolation. We examined a four independent task structure where each of the four tasks could be controlled by either a human or automation (machine). When describing potential configurations, we use (H) to denote a human in control of a task and (A) to denote automation controlling the task. Either agent could be in control of each of the four tasks. For example, HHAA would be used to denote that a human is controlling the first two tasks while the automation is controlling latter two tasks.

We computed the overall task score from each of 15 million iterations with random underlying performance distributions for the human or automation. In this mathematical validation, the underlying performance distribution for each task was distributed with a mean ranging from 0-100% with a standard deviation of 10%, capped at 0% and 100% performance. Each run of the simulated experiment was 100 time samples long. The average across-task score was calculated for situations where 1) all tasks were performed by a human, 2) all tasks were performed by automation, and 3) tasks were performed by the optimal combination of human and automation. In this lattermost condition, the agent with a higher a priori performance distribution was selected to perform each task. Figure 1 illustrates these three final outcome distributions and provided us with a benchmark the AM's ability to successfully increase overall performance with no prior assumptions regarding human and automation performance distributions. This mathematical validation demonstrates that if the algorithm is sufficiently adaptive and robust to changes in parameterization, it should lead to a significant improvement of task performance over a human or machine performing the task alone. Combined with real-world expertise (which is greater than the average of the uniform distribution used in this test) and training interventions to improve human baseline performance, implementation of the AM into a workspace can raise the human/automation baselines thus improving the HMT's performance beyond what the green HMT distribution in Figure 2 shows.

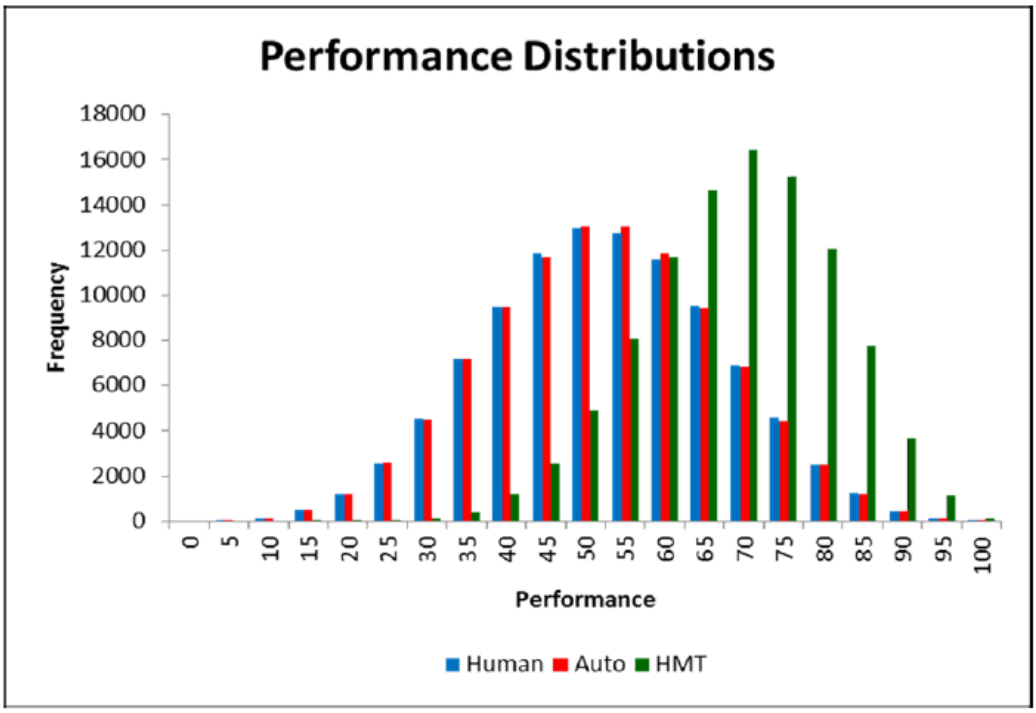

Figure 1: Histograms of performance for randomly sampled configurations of human and automation mean performance on each of the 4 tasks, as well as the distribution for the optimal HMT configuration.

The HMT distribution demonstrates consistent and robust higher performance than human or automation performance in isolation.

4.0 DEVELOPING A SIMULATED AUTONOMOUS MANAGER

After initial mathematical validation, we tested the AM simulation. The AM was designed to dynamically allocate tasks to either a human or automated agent over time, based on information about ongoing simulated task performance and indicators of human workload. Simulation studies are powerful in that they are capable of testing a larger range of plausible scenarios than would be feasible through experimentation, including possible, but improbable scenarios. This is particularly valuable for us to test extreme conditions that could break down the decision logic of the AM, allowing us to test its robustness to realistic variability. Prior to empirical evaluation, simulations serve to provide insights into mechanisms that contribute to reactions within a real-world system and can assist with guiding the specifications of research questions in an empirical test or to modify parameterization of the decision logic of tools designed to provide assistance. The overall decision logic of the AM is illustrated in Figure 2 and the following sections articulate parameters that can be modified by users.

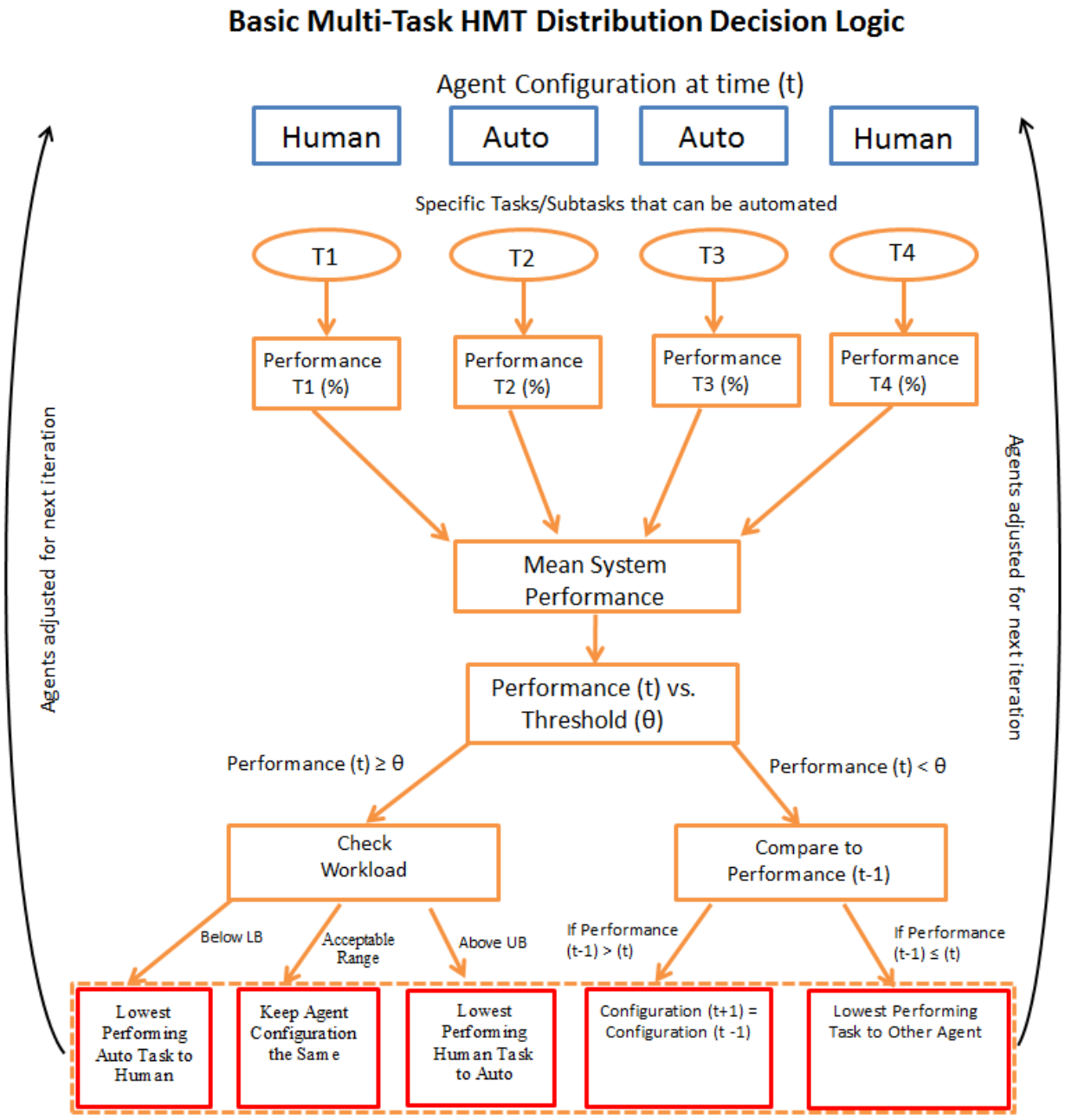

Figure 2: Decision logic of the Autonomous Manager (AM). Initial performance distributions are computed for human and automation on each of 4 tasks (either estimated or based on prior data)

These values are then simulated on each iteration and depending on HMT performance relative to a specified threshold, the current HMT performance is either checked against the previous configuration or workload is assessed. If workload is found to be above or below an acceptable range, a task is reallocated to automation or human, respectively. This logical progression represents the default configuration, but can easily be modified with user-set parameters for flexibility.

4.1 Autonomous Manager Functions and Customization

In each of the following subsections, we articulate the various user-configurable parameters of the Autonomous Manager as well as the procedural decision logic implemented by the AM to reallocate tasks dynamically between humans and automation.

4.1.1. Simulation Decision Logic

The backbone of the AM simulation is the procedural decision logic used to iterate from one configuration of Human-Machine task allocation to another adaptively and dynamically. Performance on multiple tasks was simulated continuously over time and task allocations were adjusted between human or automation based on the performance of the HMT and simulated human workload (simulated heart rate specifically). However, for the purposes of our testing environment and the priorities in surveillance, the decision logic prioritized maintaining acceptable performance over moderating workload. When the HMT performed above a minimum acceptable performance threshold, simulated physiological metrics were assessed to determine workload. If workload was above a user-specified maximum threshold, the lowest performing task by the human was taken over by the automation. Conversely, if workload was shown to be excessively low, indicating operator underwork or boredom, the lowest performing automation task was reallocated from automation to human control. When performance of the HMT was below a minimum threshold, the allocation script (AM) determined which task had the lowest performance between the two agents (human or automation) and switched from one operating agent to the other. There was a simple "memory" parameter, where the AM would reallocate the tasks to match the previous configuration if performance was substantially higher using that previous configuration, to prevent the random walk from inadvertently lowering HMT performance in successive iterations or switching rapidly between two equally suitable configurations, as this sort of rapid switching would likely confuse a human subject or lead to a loss of situation awareness. Figure 2 illustrates the flow of task redistribution logic implemented in our simulation.

4.1.2. Simulated Task Configuration

The decision logic of the AM can be implemented for a myriad of possible tasks, but was constrained for this simulation as a 4-task setup. Although the simulated AM is capable of managing up to a theoretically infinite number of simultaneous tasks, to the degree that it will eventually overwhelm computational memory, 4 tasks were selected for the analyses based on previous studies of multitasking and workload. Figure 3 illustrates a layout of 4 surveillance windows that an analyst might monitor and report events of interest on. Similar setups have been used in previous workload tasks, such as the MATB-II (Santiago-Espada, Myer, Latorella, & Comstock Jr, 2011), a well-studied multitasking environment consisting of independent tasks that has been modified for research (Blaha, Cline, & Halverson, 2015). We limited tasks in this simulated working environment to those that could be performed fully by either a human or automation and could be delegated effectively between them.

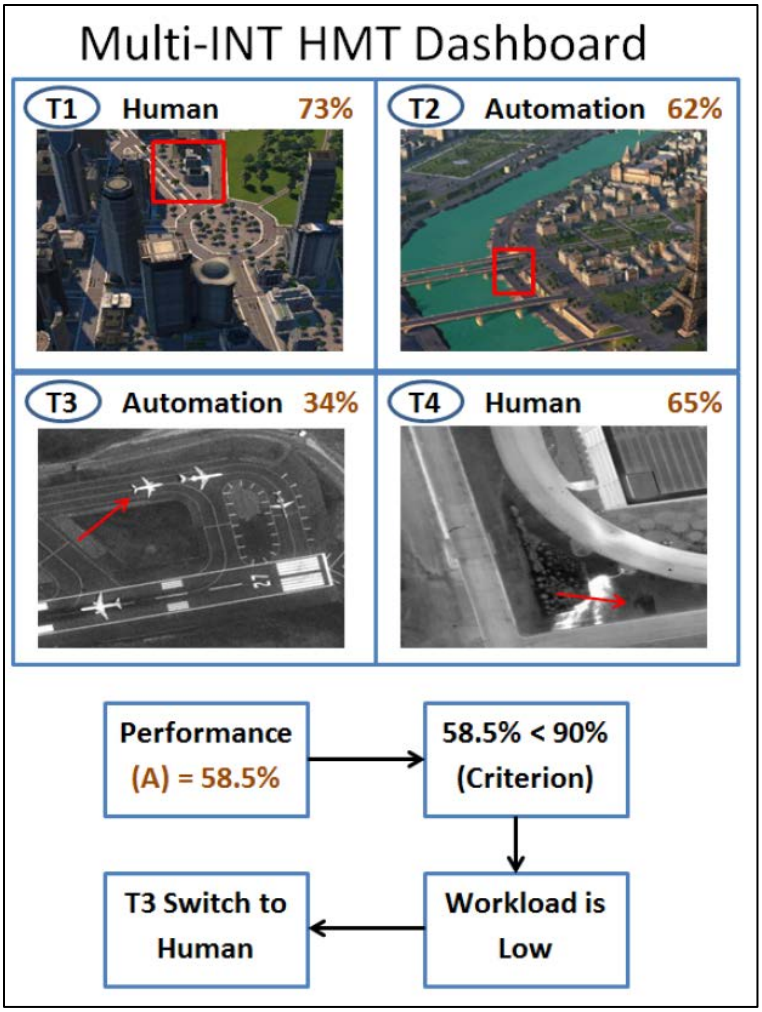

Figure 3: An illustration of a possible multi-tasking configuration consisting of multiple visual search surveillance tasks (T1, T2, T3, T4), involving either standard image/video feeds or infrared with a simplified representation of the AM's decision logic below

This figure was created with open-source images by the authors. The overall average (A) performance across all tasks is calculated. This percentage is compared to a user-set criterion, workload is assessed, and the AM makes a tasking reallocation decision.

4.1.3. Configurable AM Parameters

There are a variety of simple input parameters that a user can configure prior to running the simulated AM. User-customizable parameters fall under three primary categories: 1) experimental parameters, 2) agent (human and automation) performance parameters, and 3) human physiology parameters. In an empirical setting, the second and third categories of parameters are input into the AM as they are collected by an experimental task environment and appropriate physiological sensor, respectively.

Experiment Features. Users can set parameters relating to experiment or mission features, which are independent of human or automation performance. These parameters include

experiment duration, tempo of trials or events, and how frequently the AM will aggregate performance across trials to determine accuracy within and across tasks, based on either time (in seconds) or number of trials.

Agent Features. As with experiment parameters, agent factors are configurable simulation parameters. A priori performance distributions can be set for human and automation on each task, based on prior or estimated measures of central tendency and variance. By default, this is configured using mean and variance in a Gaussian/normal distribution. However, a user can choose to model performance using another class of distribution, such as a Poisson for count variables or an exponential for response times. In addition to specifying average accuracy, a degree of uncertainty in this mean accuracy can be set by specifying a distribution of plausible mean values on each task for each agent. By default, generated mean uncertainty is distributed using a uniform distribution, but this can be customized if a user believes that mean uncertainty should be distributed differently. For most practical applications, either a uniform or normal distribution should suffice. If there is no desire to model uncertainty, this can be omitted entirely from the simulation. Additionally, a standard deviation can be set around the performance mean(s). Finally, a user can set a minimum acceptable accuracy threshold for team performance. For example, if a user would find HMT accuracy of less than 90% averaged across all tasks to be unacceptable, the performance threshold would be set to .9. When using the AM within in situ tasking, performance values are collected and input into the AM rather than simulated. However, frequency of score aggregation is still a modifiable parameter in an empirical setting.

Physiological Features. The final set of customizable simulation parameters are metrics of human physiology pertinent to workload. In the current set of analyses, human heart rate was simulated as a proxy for workload due to the robust positive correlation between increased heart rate and increased workload (Hankins & Wilson, 1998; Luque-Casado et al., 2016). Heart rate was set to be higher on average when the human is engaging with more tasks and lower when the human is not monitoring as many tasks. These were moderately correlated to account for variability. Again, for in situ tasking this information is collected and input in real-time (or near real-time), rather than simulated.

5.0 TESTING THE A.M. OVER A LARGE PARAMETER SPACE

As an initial test of the AM, we simulated a variety of configurations to compare the performance of the AM versus an all-human baseline. For each run, we computed the mathematically optimal, baseline, and AM Parser (output of the AM's task reallocation, labeled as AM Parser in figures) performances. Mathematically optimal scores were calculated as the score that would occur if the more proficient agent maintained continuous control of tasking. This is the aggregated performance that would be yielded if the AM were omniscient about both agents' potential scores at each time point and how these dynamics would change with different task configurations. Baseline was defined as the score that would be obtained if a human were performing all four tasks since this is how tasks are currently distributed in surveillance environments. Although there are other potential baselines that could have been generated, such as randomized parsing, human performance alone serves as the most accurate proxy for task allocation under current capabilities. Presently automation has not been integrated into performing higher-level cognitive tasks beyond repetitive tasking and "busy work" such as copying and pasting text. However, as automation for complex tasking is under development, it is important to concurrently develop a means of managing HMTs before full-scale implementation of automated solutions. The mathematically optimal distribution along with the baseline distribution allowed us to determine where the AM Parser results fall between current configurations and optimal redistribution of tasks.

5.1 Simulation Results

Our initial simulations were run across a wide span of potential performance and realistic heart rate parameterizations. We divided these simulations into two categories: 1) performance distributions remaining consistent over time, and 2) performance distribution for the human changing over time to reflect inherent nonstationarities due to workload and fatigue fluctuations.

5.1.1. Determining Degree of Improvement in a Simulated Multitasking Environment

The human, automation, and HMT distributions were constructed from 10,000 simulations of a four-independent task configuration. Mathematically optimal (labeled Optimal in plots) HMT performance values were calculated for each run as well as average human-only scores to generate the baseline distributions (see Figure 1 in Section 3). In Figure 4, the distribution of AM Parser scores shows substantial improvement over baseline and nearly approximates the mathematically optimal distribution.

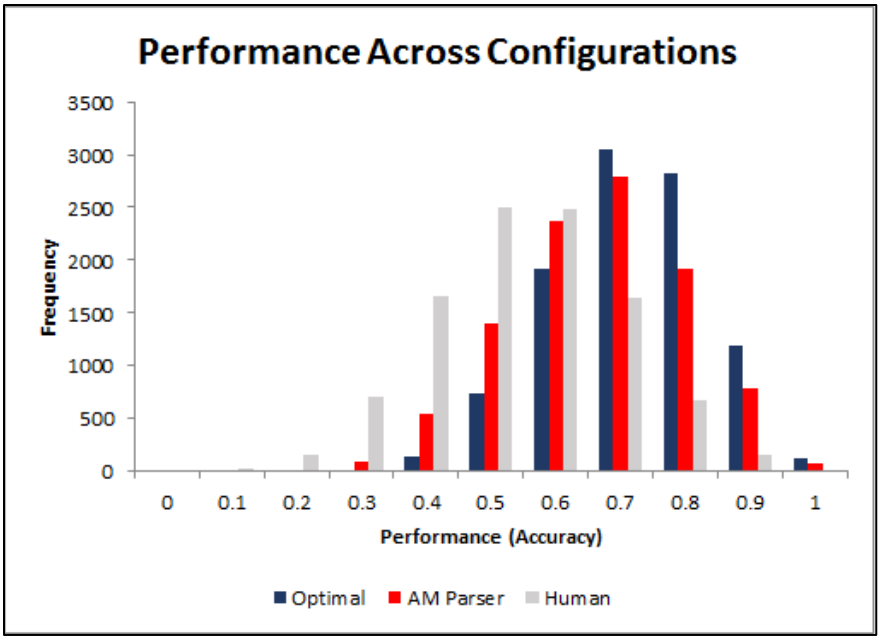

Figure 4: Distributions from 10,000 iterations of the AM simulation

The gray distribution indicates the baseline performance (human performing all tasks). The blue distribution indicates the performance from the mathematically optimal allocation of tasks within the HMT, given each of their respective performance constraints. The red distribution is the resulting HMT performance based on the Autonomous Manager's parsing.

With the AM, there was an overall improvement of M = 11.36%, SD = 8.64%, and ranged from ≈ -10% performance improvement (lower than baseline) to an ≈ 40% improvement over baseline (see Figure 5 for the improvement distribution). Using the AM, paired samples t-tests were run comparing the AM Parser and optimal performance distributions, and comparing the AM Parser and baseline performance. AM Parser performance was significantly higher (M = 66.33%, SD = 13.59%) compared to baseline performance (M = 49.96%, SD = 14.58%), $t(9999) = 85.10$, $p < .001$, $d = 1.16$.

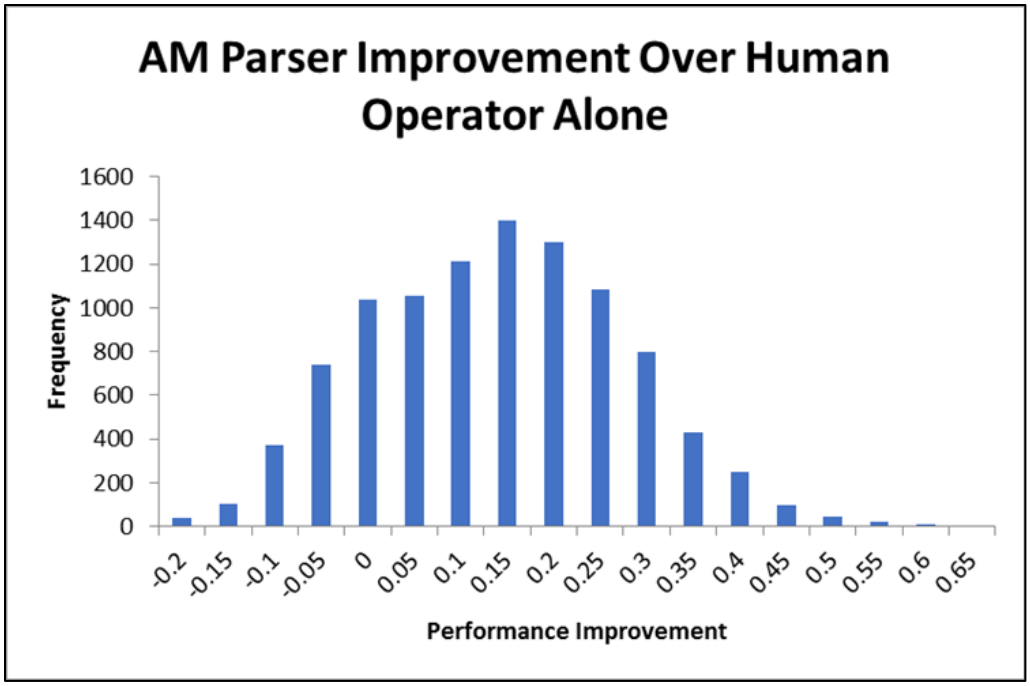

Figure 5: Distribution of overall (average) improvement between the AM Parser and Human-only baseline for each run of the simulation

The typical range of improvement over baseline was between 5-25%.

The superior performance of AM Parser over baseline indicates that the use of the AM led to significantly higher overall scores than one would expect from the human performing the tasks alone. However, there was also a significant difference between performance with the AM and the mathematically optimal performance (M = 66.79, SD = 11.95%), $t(9999) = 88.07$, $p < .001$, $d = .036$, indicating that there is likely still room for improvement of the decision logic. This is unsurprising given the simplicity of the simulation logic and use of static thresholds. However, the AM results were closer to the optimal distribution than the baseline distribution and the effect size for the comparison between the AM Parser and mathematically optimal was small, especially compared to the AM Parser vs. baseline performance. The significant difference between the AM Parser and the optimal performance is likely due to the large number of iterations conducted for the analysis (10,000 iterations). The Cohen's d effect size for the comparison between the baseline and AM Parser however is relatively high and most likely reflects a genuine difference in scores, indicating that there is room for improvements as the AM is refined and calibrated to novel tasking environments.

5.1.2. Robustness to Nonstationary Performance

Within a practical working environment, human and automation performance will likely be nonstationary as well as stochastic, as indicated by both increasing/decreasing drifting mean performance over time and short-term variability of performance, respectively. For example, one might expect fatigue to lead to a gradual drift or decrement of performance with a potential

increase in performance after a rest break so that fatigue can be mitigated. We simulated three configurations of changing performance to test the AM for robustness to nonstationarity. In the first test, mid-way through the simulation, we changed the mean human and automation performance for each of the four tasks. For the second test, we increased the frequency of these changes to modify the human and automation performance at each quarter, and for the third test we changed performance mean parameters six times with random periodicity, rather than by a specific temporal interval. By randomizing periodicity, this meant that the random walk would not always have time to recover if parameterizations changed within a short period of time. All changes were made abruptly, rather than gradually ramping up or down, in contrast to the function for real-world fatigue. This provided the AM with a "worst case" scenario of wildly fluctuating performance parameters, which would require quick adaptation by the AM. Figure 6 illustrates an example of how performance might fluctuate in a single run, under each of these conditions. For ease of visual interpretation, only two task performances are plotted over time, rather than four. If the AM is robust to this changing mean performance, this will demonstrate the power and flexibility of the decision logic regardless of variability and perturbations, and demonstrate the value of the AM for real-world applications.

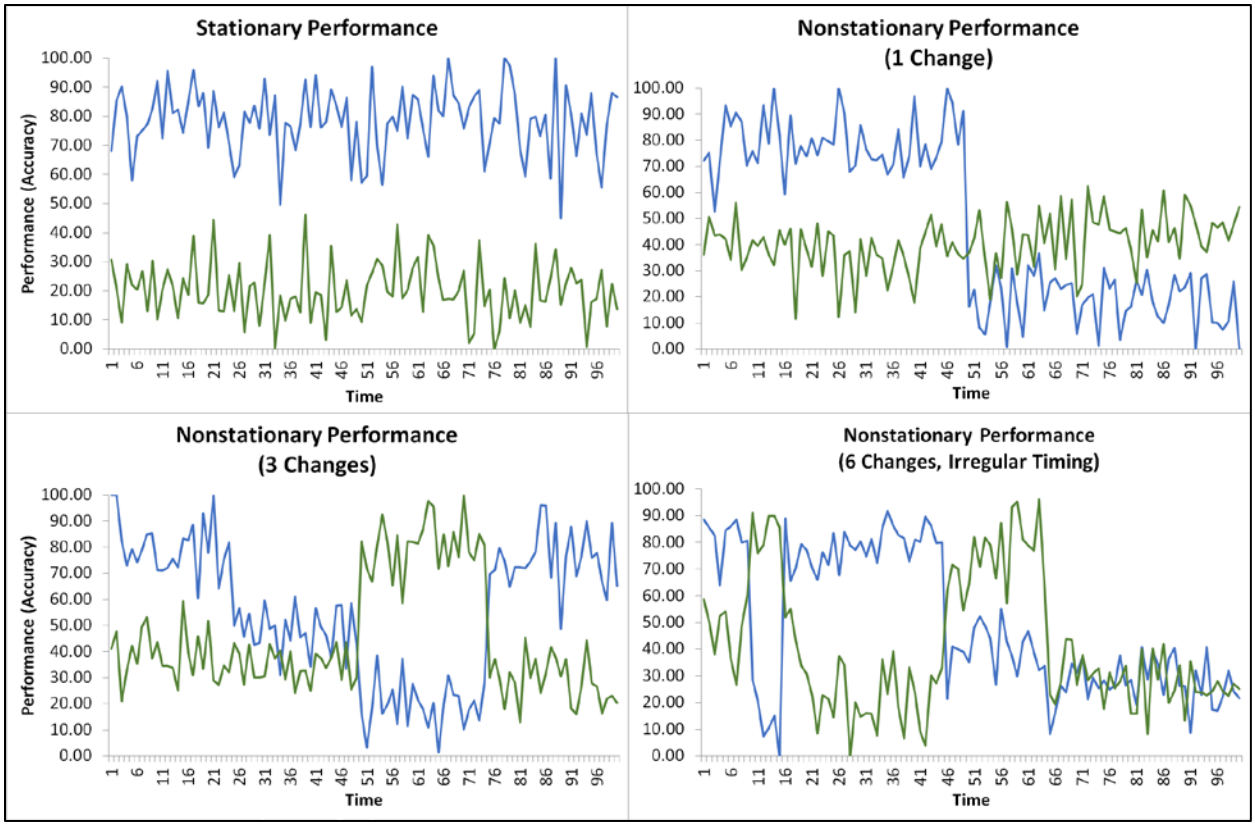

Figure 6: The effect of nonstationarity on performance of two hypothetical tasks. In the upper left panel, the performance data is simulated to stochastically vary, but remain stationary

However, in the other three panels (Upper Right: 1 Change, Bottom left: 3 Changes, Bottom Right: 6 Random-Interval Changes), there are multiple deflections where greater shifts in performance occur based on changes in mean performance. In all four panels, the standard deviation of performance is equal to 10%.

As with the stationary condition, 10,000 simulations were conducted for each of the three nonstationary conditions, with an initially randomized performance parameterization. For the first condition, mid-way through each simulation, the parameterization for both human and automation was randomly changed to a new mean performance value (retaining the same standard deviation of 10). As with the stationary distributions, for the single deflection nonstationary case there was a significant difference between AM parsed and mathematically optimal performance (see Table 1 for summary statistics for all nonstationary distributions).

Table 1: Results of Paired T-tests for Nonstationary Conditions (Cohen's d for Effect Sizes)

Nonstationary Summary Statistics					
Condition	**Optimal_Mean (SD)**	**AMParser_Mean (SD)**	**t**	**p**	**Effect Size**
Nonstationary (1 Change)	66.81% (11.92%)	61.29% (9.45%)	58.16	<.001	0.513
Nonstationary (3 Changes)	66.97% (11.99%)	60.85% (6.64%)	56.07	<.001	0.631
Nonstationary (6 Changes, RI)	67.04% (11.91%)	60.01% (5.63%)	57.19	<.001	0.755
Condition	**Baseline_Mean (SD)**	**AMParser_Mean (SD)**	**t**	**p**	**Effect Size**
Nonstationary (1 Change)	50.09% (14.53%)	61.29% (9.45%)	80.67	<.001	0.914
Nonstationary (3 Changes)	50.06% (14.62%)	60.85% (6.64%)	74.59	<.001	0.948
Nonstationary (6 Changes, RI)	49.75% (14.77%)	60.01% (5.63%)	67.63	<.001	0.918

The table of results for each of the nonstationary conditions illustrates that there were significant mean differences between the baseline and AM Parser with relatively large effect sizes. This is consistent with the outcome of the simulation using stationary HMT performance. As with the stationary condition, there were also significant differences between the HMT score from the AM Parser compared with the optimal potential score, with more moderate effect sizes. Figure 7 provides an illustration of the outcome performance distributions for each of the nonstationary conditions.

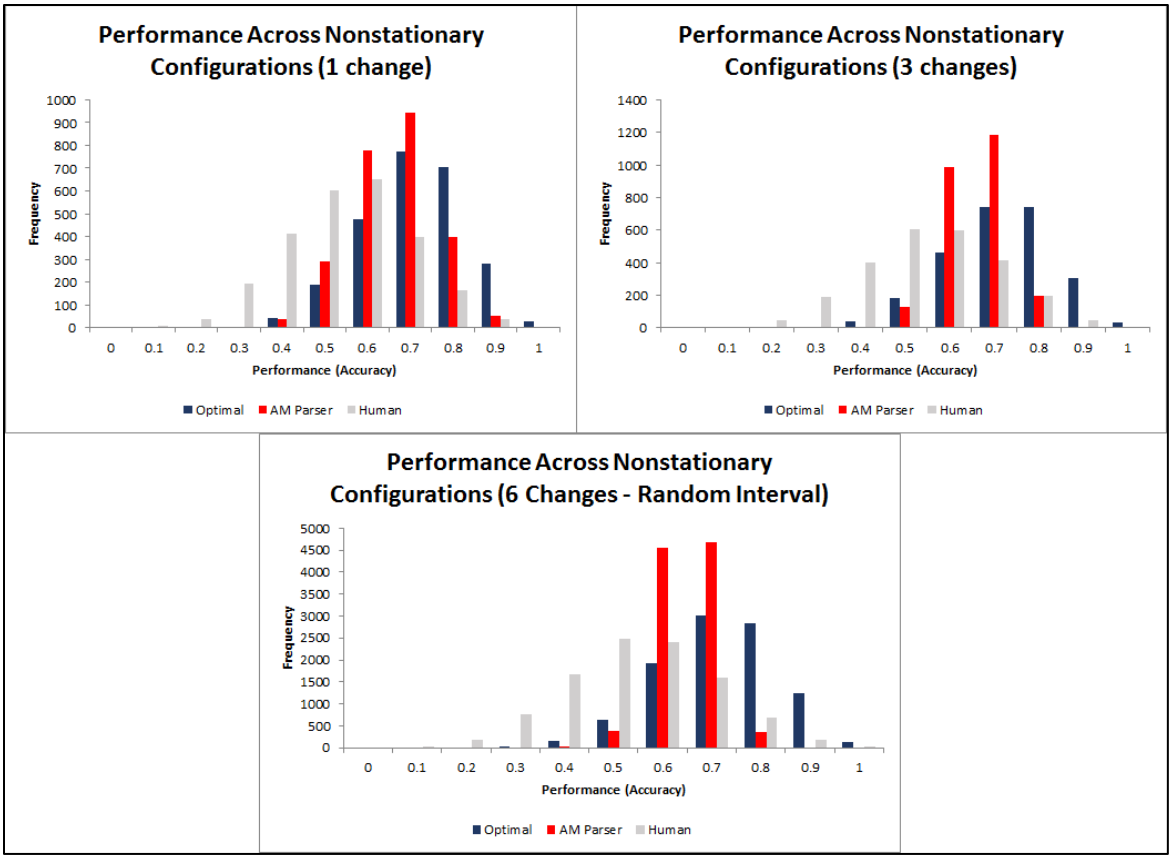

Figure 7: Distributions from 10,000 random iterations of the AM simulation with a change in human and automation performance distributions occurring mid-way through the simulation (top left), three times during the simulation quarterly (top right), or six times at random intervals (bottom graph)

As with the distributions for stationary performance, there is a significant difference between baseline performance and performance from the AM Parser and a significant difference between the AM Parser's performance and the mathematically optimal performance.

Figure 8 illustrates the degree of improvement for each nonstationary condition when parsed by the AM over baseline. For both stationary and single deflection nonstationary conditions, improvement over baseline (human performing task alone) was calculated. The mean performance improvement for the stationary simulations from Section 5.1.1 (M = 11.47%, SD = 13.67%), was not significantly different from the nonstationary simulations' performance improvement (M = 11.20%, SD = 13.89%), $t(9999) = 1.40$, $p = .162$, $d = .020$.

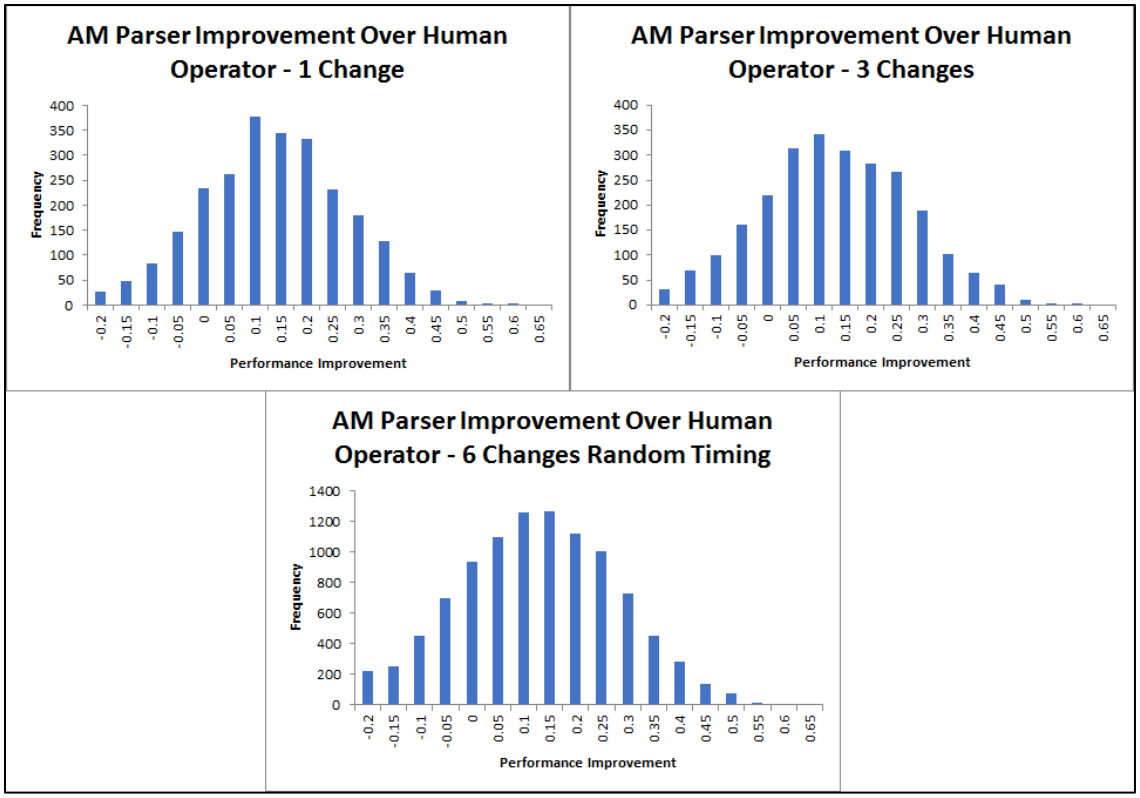

Figure 8: The degree of improvement over baseline performance for all three nonstationary conditions is extremely similar to the distribution from the stationary condition, with less than a single percent difference in mean improvement between all conditions

Further simulations were run with three changes of mean performance across all tasks, occurring at equal intervals, effectively leading to a different mean performance for each quarter of a single simulated experiment. Under these circumstances, the thrice changing nonstationary condition (M = 10.79%, SD = 14.47%), yielded significantly lower performance improvement than the stationary condition (M = 11.47%, SD = 13.67%), $t(9999) = 3.43$, $p = .011$, $d = .048$. This is not entirely surprising since with increasing frequency of performance fluctuations, the more adjustments the random walk must make to converge to a higher performing task reallocation. Although the mean difference is robust enough to lead to a statistically significant difference, the more important factor in an applied setting is whether the mean difference is practically meaningful. With a performance decrement of only .68%, in most real-world environments this would not be a meaningful difference. The low effect size provides additional credence that the significant difference is likely due to large sample size rather than a robust effect. The final analysis allowed mean performance to change six times across a single simulated experiment at random intervals, which would prevent a guaranteed recovery time for the random walk process after perturbation. Under this condition, the nonstationary performance improvement (M = 10.26%, SD = 15.16%) again was significantly lower than the stationary performance (M =

11.47%, *SD* = 13.67%), *t*(9999) = 5.97, *p* < .001, *d* = .084. However, with an increased number of performance fluctuations occurring at less predictable times, this difference is unsurprising. Again, however, a mean difference of only 1.22% is not practically significant and the low effect size indicates that this significant difference is likely spurious, due primarily to the large number of iterations in the simulation.

Table 2 provides results from a 1-Way ANOVA comparing degree of improvement over baseline across all four stationarity conditions. The omnibus test was found to be significant, *F*(3, 39996) = 13.79, *p* < .001. However, the only significant pairwise comparison according to a Tukey HSD posthoc test was the mean difference between the Stationary condition and the Nonstationary condition with 6 random-interval changes.

Table 2: ANOVA Comparing Stationary vs. Nonstationary Conditions

Groups	*Average*	*Variance*				
Stationary	0.1147	0.0187				
Non-Stationary (1 Change)	0.1120	0.0193				
Non-Stationary (3 Changes)	0.1079	0.0209				
Non-Stationary (6 Intermittent Changes)	0.1025	0.0230				
Source of Variation	***SS***	***df***	***MS***	***F***	***P-value***	**Eta Sq.**
Between Groups	0.847	3	0.2824	13.790	< 0.001	0.001
Within Groups	818.896	39996	0.0205			
Total	819.743	39999				
Tukey HSD Test	***P-value***					
Stationary vs. NS_1	p > .05					
Stationary vs. NS_3	p > .05					
Stationary vs. NS_6	p < .01					
NS_1 vs. NS_3	p > .05					
NS_1 vs. NS_6	p > .05					
NS_3 vs. NS_6	p > .05					

Importantly, the effect size for the omnibus test is extremely small, indicating that the degree of stationarity poorly explains the variability of performance improvement. This means that the distributions of performance improvement are similar regardless of how frequently the AM must make adjustments due to nonstationary performance (see Figure 9 for means of each condition). The AM shows practical robustness to dynamic performance conditions, even in the tested "worst case" scenarios.

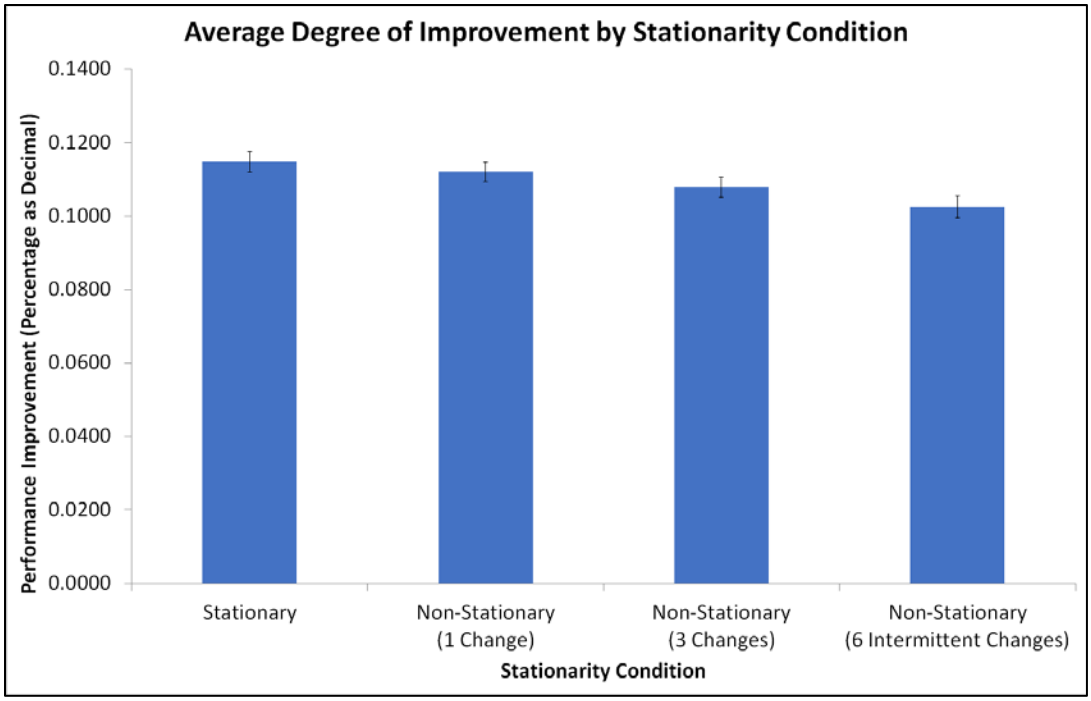

Figure 9: Mean performance improvement for Stationary vs each Nonstationary condition
The only significant difference is between Stationary and the Six-Intermittent Change condition, but at approximately 1% performance disparity, this statistically significant difference is not practically distinct for most real-world environments.

6.0 SIMULATED AM FOR PRACTICAL IMPLEMENTATION AND TESTING

The initial tests of the simulated Autonomous Manager demonstrated the value of an intelligent decision logic incorporating performance and physiological correlates of human workload to increase team performance and prevent overwork. In an empirical task or a working environment with established workflows, the AM can be incorporated in real-time. However, in certain applied environments, there may be multiple candidate tasking configurations. This may include fewer or greater than the four tasks tested in our simulations. Additionally, the optimal tasking configuration may be unknown. As an example of testing multiple potential task combinations, one can imagine a manager who wants a surveillance worker to perform three simultaneous tasks with an automated teammate. There are eight possible tasks to choose from in this hypothetical working environment. This means there are 56 total task configurations that are possible. The manager will only employ the AM for tasking configurations where it yields the greatest degree of improvement over the human performing the tasks alone, since it isn't worth the cost of implementing automation in instances where there isn't a substantial improvement. This is a necessary adjustment when automation is expensive to develop or is not feasible to employ for all tasks. It may not be tractable to test every single one of these task combinations empirically using the AM in real-time. However, as the developed simulation uses some of the same decision logic, it can be utilized prior to full scale implementation to test the degree of marginal benefit from using the AM for different tasks, provided there is some a priori estimate of performance or previously collected data regarding typical task performance for both a human operator and automated agent.

We tested two hypothetical use-case scenarios calibrated to a surveillance environment and typical multitasking setup:

1) Testing Multiple Combinations of Task Subsets
2) Human and Automation Expertise

6.1 Use-Case 1: Testing Multiple Task Subset Combinations

For the first use-case, we tested a similar scenario to the hypothetical manager scenario discussed in Section 6.0. Specifically, we tested a plausible real-world scenario where a manager might have the choice of two or three tasking configurations for a HMT to perform. There are six hypothetical candidate tasks (Vigilance, Visual Search, Categorization, Decision Making, Mental Rotation, and Cueing), but the HMT will only perform four of them simultaneously. The top table of Table 3 provides a description of the hypothetical task and performance distributions, as well as known workload levels and fatigue effects from hypothetical a priori testing of each task individually. To account for difficulty, there was a higher penalty to workload when humans were engaged with a given task (Heart Rate penalty of 0-3 Beats per Minute for Easy tasks and 0-6 Beats per Minute for Difficult tasks). These modifications are somewhat arbitrary without real-world data, but were designed to provide an explicit change to workload while being minimally influential compared to performance. The degree of change in heart rate is small to

reflect minute changes that occur within a seated stationary tasks where the human is not actively engaging in high cardiovascular activity, as would be seen in tasks involving movement. However, given that these tasks are cognitively difficult rather than physically demanding (all are sedentary computer tasks), these are not unreasonable values since the maximum increase is only 24 Beats per Minute. Fatigue was denoted via a gradual decrease in mean performance at the same time points as the 3-Change Nonstationary condition from Section 5.1.2, with Slight fatigue yielding a 0-5% decrease in mean performance at each change and Moderate yielding a 0-10% decrease, uniformly distributed. To maintain tractability, we presumed that the hypothetical supervisor narrowed the tasking configurations to the three most useful (Table 3, bottom table), but wants to determine which one will yield the highest HMT performance improvement with adequate autonomous management to maximize the AM's benefit to performance. For each of these three tasking configurations, 10,000 iterations with 100 time bins each were run, with an 85% minimum performance threshold.

Table 3: (top table) Simulated Use-Case Parameterizations for 6 Tasks with Means (as percentages), Standard Deviations, Workload (2 levels), and Fatigue (2 levels) (bottom table) Task Configurations Considered for Use-Case 1

Potential Task	Human Performance Mean (SD)	Automation Performance Mean (SD)	Workload Level	Fatigue Effects
Vigilance	90 (5)	85 (10)	Easy	Slight
Visual Search	65 (10)	90 (5)	Easy	Moderate
Categorization Task	60 (5)	80 (15)	Easy	Slight
Decision Making	80 (15)	65 (10)	Difficult	Moderate
Mental Rotation Task	85 (10)	70 (15)	Difficult	Moderate
Cueing Task	70 (15)	60 (5)	Difficult	Slight

Combination	Task 1	Task 2	Task 3	Task 4
A	Vigilance	Visual Search	Categorization	Decision Making
B	Visual Search	Mental Rotation	Decision Making	Cueing Task
C	Vigilance	Visual Search	Cueing Task	Categorization

6.1.1. Use-Case 1 Results

The results of the simulation indicated significant differences in performance improvement between the human baseline performance and HMT performance when tasks were allocated by the AM for all three Configurations (A, B, and C, see Table 4).

Table 4: Multiple Task Configurations – HMT Improvement

	HMT > Baseline	Optimal > HMT	T	P
Combination A	3.32%	1.33%	78.15	< .001
Combination B	2.26%	1.08%	273.00	< .001
Combination C	6.20%	1.30%	86.26	< .001

However, the degree of improvement also varied significantly between conditions. Figure 10 provides a comparison of the mean performance values of the mathematically optimal score, the human baseline score, and the HMT score. The greatest degree of improvement using the AM parsed HMT was found with task Combination C and the least substantial improvement was found using task Combination B.

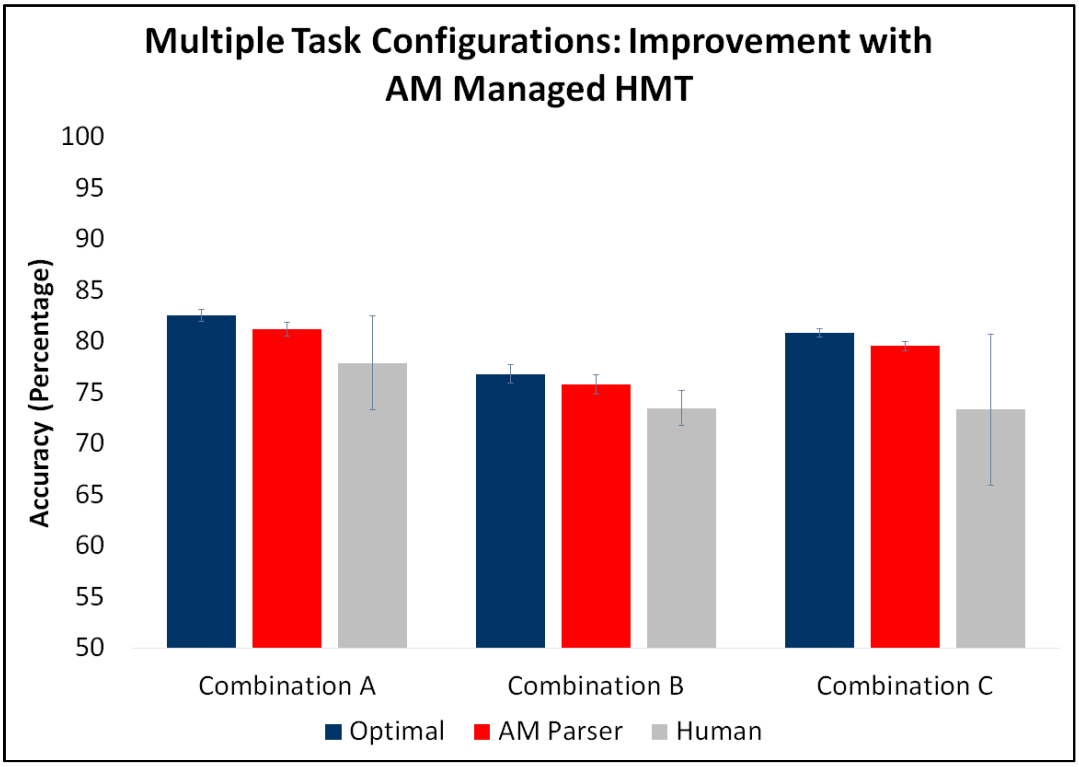

Figure 10: For all three conditions, there was a significant improvement in performance when the HMT was controlled by the AM

However, the greatest improvement was found in task Combination C (Vigilance, Visual Search, Cueing Task, and Categorization), while the smallest improvement was found in Combination B (Visual Search, Mental Rotation, Decision Making, and Cueing Task).

Interestingly, although perhaps not surprisingly, the mean number of tasks performed by the human teammate was strongly negatively correlated with performance improvement (all correlations stronger than -.52, $p < .001$). The human is tasked with performing the greatest number of tasks in Combination B and the lowest number in Combination C (see Figure 11).

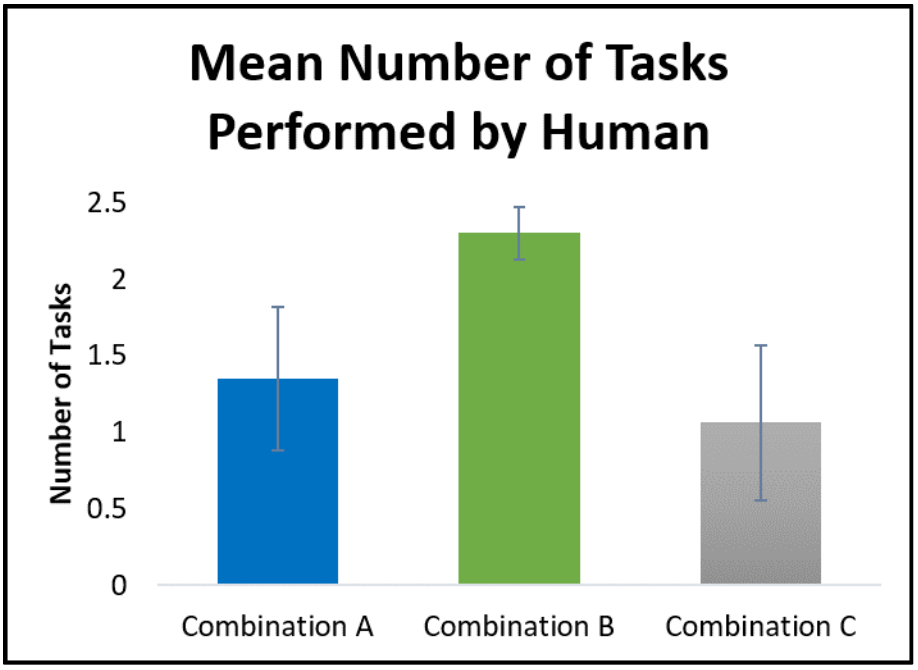

Figure 11: On average, with task Combination B, the human teammate was tasked with controlling a larger number of tasks than in the other two conditions

Number of tasks performed by the human was moderately negatively correlated with the degree of performance improvement by leveraging the AM over the baseline performance.

If a manager wanted to leverage the AM for the greatest performance improvement, they would probably choose task Combination C. For all three conditions, there was a significant, but not practical (≈ 1%) difference between HMT and optimal performance. There was no significant difference in mean heart rate between the three task configurations, as it was maintained at a healthy 66-87 BPM. This is due to the AM's decision logic that strives to maintain a healthy workload in addition to adequate HMT performance.

6.2 Use-Case 2: Effects of Expertise

Another potentially valuable use-case is to determine the AM's effectiveness to parse tasks between expert agents, novice agents, and agents where there is a mismatch between human versus automation expertise. The simulation can be leveraged to determine if under these conditions, the AM provides excellent or only marginal benefits. Expertise was fully crossed between agents in a 2x2 design. The generating distribution for mean performance on all tasks was (M = 85%, SD = 10%) for experts and (M = 65%, SD = 25%) for novices.

6.2.1. Use-Case 2 Results

Findings from the second use-case pertaining to expertise were as expected. There was a ceiling effect when both agents had high expertise and a floor effect when they both had low expertise. This demonstrates that when both perform around threshold performance, performance will be maintained at exactly that threshold value. When both agents' performance is below threshold,

even the optimal parsing would be unable to overcome these inadequacies. There was a significant difference between the HMT performance and human-only baseline for the both expert condition, $t(999) = 4.08$, $p < .001$, and the no expert condition, $t(999) = 3.06$, $p < .001$, but with mean differences of 0.011% and 0.014% respectively, this is not practically meaningful. There was a significantly higher baseline in the human-expert only condition compared to the HMT, $t(999) = 329.95$, $p < .001$, but again at < 1% mean difference, this is not practical. However, this does indicate that in situations where the human dramatically outperforms the automation, there is no need for teaming, other than to mitigate high workload. HMT performance was expectedly higher than human baseline performance when the automation was the expert agent, $t(999) = 1012.75$, $p < .001$, but interesting, HMT performance was significantly superior to the automation's solo performance as well, $t(999) = 332.71$, $p < .001$. However, this was by a mean difference of < 1%, denoting that the AM provides whatever small benefit it can to increase overall performance by leveraging the combined efforts of both agents in the HMT. There were no significant differences between HMT performance and the optimal score when both teammates had equal expertise. HMT was significantly, but not practically (mean difference < 1% in both cases) lower than optimal when there were asymmetries in partner expertise. The number of tasks maintained by the human partner was also interesting to compare based on partner expertise. When expertise was asymmetrical, the number of tasks for the human either converged to four (all tasks) when the human was the expert, or to zero when the automation was the expert (see Figure 12). When both teammates displayed equal expertise, the number of tasks converged to a 2/2 split of tasks between the human and automation, but with a substantial degree of variability. This variability was highest when both agents were experts, compared to when both were novices.

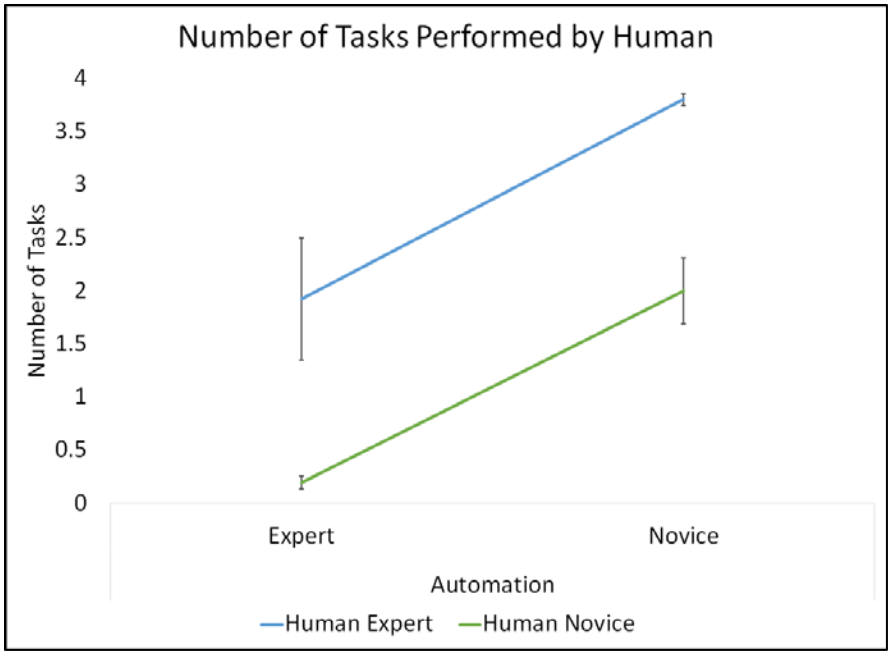

Figure 12: The number of tasks performed by the human varied predictably as a function of expertise

When there was an imbalance of agent expertise, the number of tasks asymptotically approached either all or none of the tasks. When both agents were either experts or novices, the AM balanced the number of tasks for the human at ≈ 2 tasks, but with a fairly high degree of variability.

7.0 CONCLUSIONS AND DISCUSSION OF SIMULATION RESULTS

These simulation studies demonstrate the potential usefulness of an Autonomous Manager (AM) that can adaptively redistribute tasks within HMTs in dynamic environments via a range of simulations. It can also be used to test a variety of possible performance scenarios, including unlikely scenarios that might not occur in many empirical tests prior to real-world implementation. This provides not only the financial benefit of reduced cost, but elucidates the range of potential benefit of implementing adaptive automation. This can be particularly helpful in environments where the cost of running a real-world task is expensive or situations where there are multiple plausible outcomes. Additionally, this simulation has been used as a step toward calibrating and finalizing the decision logic of the AM, which has been empirically tested in a realistic multitasking visual-search paradigm (Frame, et al., 2019). Continuous improvements to the simulation allow for greater flexibility in real-world implementation and vice versa, meaning the AM is consistently being refined and improved. As the AM currently functions, it is reminiscent of the adaptive automation recommended by Parasuraman and colleagues (2000) and Berka and colleagues (2007), both of which asserted that changes in tasking should be adaptive and informed by some kind of human physiology variables. The current infrastructure of the AM is capable of either simulating or having physiology data input to inform how tasks should be dynamically distributed based on HMT performance and cognitive workload.

This, as well as continuing practical development of the AM, has the potential to redefine some of the roles within applied multitasking ISR environments, including adding a managerial component to HMTs, with possibilities for expansion into teams of HMTs. To expand on the current decision logic, we are in the process of incorporating advanced models for measuring physiological workload and concurrent input from Cognitive Metrics Profiling (CMP, Gray, Schoelles, & Myers, 2005) to better estimate appropriate high and low workload threshold. This will allow us to also determine which cognitive modalities are being taxed during each task, which will be dependent on the tasking environment. Understanding the modalities that are currently being overworked will assist with the adaptive logic for task switching, allowing a task to be given to automation that requires the human's overtaxed cognitive resource. Additionally, depending on the structure of the multiple concurrent tasks, the AM should be capable of more successfully diagnosing errors between tasks with multiple dependences. For example, in a multitasking environment, there may be information from one task that is used to inform a later task. Poor performance on the second task could be due to errors on that task itself or due to errors from the prior, dependent task. Proper intervention from the AM would require reallocation of the task causing errors for the human, to the automated teammate. Although this scenario requires a far more complex decision logic for accurate performance diagnosis, we are in the process of refining the AM's decision logic to yield greater adaptability in tasking environments with multiple interdependencies.

Overall, thorough simulations provided evidence for significant performance improvements using the AM to dynamically parse tasks, with an average improvement of around 11%. In approximately one quarter of the simulations, there was a greater than 25% performance improvement, indicating that within certain HMTs there is tremendous potential for performance improvement. Additionally, the simulation's decision logic was robust to nonstationary,

indicating that it can continue to have value and provide performance improvement even in a chaotic or noisy practical environment.

Despite the exciting possibilities for improvement using the AM's decision logic, there is ample possibility for future development and increased sophistication in this logic. Despite the simplicity of the static thresholds implemented and lack of a complex physiological model, the AM still led to a considerable performance improvement. However, further instantiations of the AM will incorporate more advanced models of workload, such as integrating eye tracking metrics, cardiac variability, and subjective perceived workload to determine appropriate workload bounds.

In practice, even a maximally calibrated AM is not the only tool that should be implemented in surveillance or other real-world environments, but rather it serves as a compliment to other augmentation tools and training. Human performance must be kept at an adequate baseline to support inclusion within a HMT, which means that we must continue to rely upon adequate training and learning of tasks. The benefit afforded by the AM is the capacity to maintain excellent performance on more tasks simultaneously to accommodate the increasing demands of the working environment.

It should be noted that our tests of the AM's dynamic task reallocation put the AM at a disadvantage compared to the typical tasking environment. Under normal circumstances, it is unlikely that nonstationarity of performance would lead to dramatic fluctuations on all tasks for both the human and automation. Typically, one would expect automation performance to stay relatively consistent with only human performance fluctuating at a slow and steady drift, and most likely only on one or two tasks at a time based on the perceptual and cognitive resources shared by the tasks. By testing robustness of the AM with all tasks changing for both agents, we provided a demonstration of robustness of the AM under a "worst case scenario", where the random walk must completely re-establish the optimal configuration over multiple steps. The fact that he AM was still able to demonstrate value under these simulated conditions is a testament to the incredible potential of adaptive automation in real tasking.

8.0 FUTURE DIRECTIONS

Parasuraman et al. (1996) provided an early demonstration that adaptive allocation of tasking can mitigate the out of the loop problem in HMT and improve the ability for humans to detect automation failures. However, humans performed poorly at detecting automation failures, even with adaptive automation, when they were simultaneously engaged with other tasks (Parasuraman, Mouloua, Molloy, & Hilburn, 1993). This is largely due to a lack of sufficient cognitive resources to both manage one's own tasking and additionally monitoring the automation's tasking. Research on interruptions have demonstrated that a similar dual task occurs when a person must decide whether to interrupt a task they are engaged in to switch to another task (Katidioti, Borst, van Vugt, & Taatgen, 2016), which is alleviated somewhat by allowing an external automated agent to determine when to switch tasks. Using a similar logic, we have continued to develop our AM parser to be applied to multitasking environments involving multiple dependencies between tasks. We have leveraged the lessons learned from this series of simulation studies to develop an AM that is integrated into a multitasking environment involving four simultaneous visual search tasks (Frame, et al., 2019). Although our initial tests have been developed with a general orientation to surveillance-pertinent task configurations, the logic of the AM is more domain general and we plan to test this in a variety of real-world task environments across multiple domains. This is an appropriate tool that can be calibrated for nearly any multitasking environment, even those where there are interdependencies between tasks, with only minor modifications to the decision logic. We plan to expand the application of the AM to task environments with multiple dependencies, teams of HMTs, and environments where the number of tasks may vary as a function of performance or workload. Currently, we are in the process of preparing a series of empirical studies examining the potential for the AM to influence task feedback and scaling task difficulty.

9.0 REFERENCES

Berka, C., Levendowski, D.J., Lumicao, M.N., Yau, A., Davis, G., Zivokvic, V.T., & Craven, P.L. (2007). EEG correlates of task engagement and mental workload in vigilance, learning, and memory tasks. *Aviation, space and environmental medicine, 78*(5), B231-244.

Billings, C.E. (2018). *Aviation automation: The search for a human-centered approach.* CRC Press.

Blaha, L.M., Cline, J., & Halverson, T. (2015). Modeling the workload capacity of visual multitasking. In *Proceedings of the international conference on cognitive modeling* (pp. 37-38).

Buettner, R. (2013) Cognitive workload of humans using artificial intelligence systems: towards objective measurement applying eye-tracking technology. In *Annual conference on artificial intelligence* (pp. 37-48).

Byrne, E.A., & Parasuraman, R. (1996). Psychophysiology and adaptive automation. *Biological psychology*, *42*(3), 249-268.

Diekfuss, J.A., Ward, P., & Raisbeck, L.D. (2017). Attention, workload, and performance: A dual-task simulated shooting study. *International Journal of Sport and Exercise Psychology*, *15*(4), 423-437.

Evans, J. S. B. (2008). Dual-processing accounts of reasoning, judgment, and social cognition. *Annual Review of Psychology*, *59*, 255-278.

Feigh, K.M., Dorneich, M.C., & Hayes, C.C. (2012). Toward a characterization of adaptive systems: A framework for researchers and system designers. *Human Factors*, *54*(6), 1008-1024.

Frame, M.E., Boydstun, A.S., Maresca, A.M, & Lopez, J.S. (2019). Development of an autonomous manager for dyadic human-machine teams in an applied surveillance environment. In *Proceedings of the international conference on intelligent human systems integration* (pp. 706-711). Springer, Cham.

Gable, T.M., Kun, A.L., Walker, B.N., & Winton, R.J. (2015). Comparing heart rate and pupil size as objective measures of workload in the driving context: Initial look. In *Adjunct proceedings of the 7th international conference on automotive user interfaces and interactive vehicular applications* (pp. 20-25).

Gray, W. D., Schoelles, M., & Myers, C. W. (2005). Profile before optimizing: A cognitive metrics approach to workload analysis. CHI Extended Abstracts on Human Factors in Computing Systems (4), 2, pp 1411-1414.

Hancock, P., Williams, G., & Manning, C. (1995). Influence of task demand characteristics on workload and performance. *The International Journal of Aviation Psychology*, *5*(1), 63-86.

Hankins, T.C., & Wilson, G.F. (1998). A comparison of heart rate, eye activity, EEG, and subjective measures of pilot mental workload during flight. *Aviation, space, and environmental medicine*, *69*(4), 360-367.

Hershey, P., Wang, M.C., Graham, C., Davidson, S., Sica, M., & Dudash, J. (2012, October). A policy-based approach to automated data reduction for intelligence, surveillance, and reconnaissance systems. In *MILCOM 2012-2012 IEEE Military Communications Conference* (pp. 1-6). IEEE.

Hershey, P., & Wang, M.C. (2013, April). Composable, distributed system to derive actionable mission information from intelligence, surveillance, and reconnaissance (ISR) data. In *2013 IEEE International Systems Conference (SysCon)* (pp. 460-467). IEEE.

Hoover, A., Singh, A., Fishel-Brown, S., & Muth, E. (2012). Real-time detection of workload changes using heart rate variability. *Biomedical Signal Processing and Control, 7*(4), 333-341.

Irvine, J.M., & Nelson, E. (2009, May). Image quality and performance modeling for automated target detection. In *Automatic Target Recognition XIX (*vol. 7335, p. 73350L). International Society for Optics and Photonics.

Kaber, D.B., & Endsley, M.R. (2004). The effects of level of automation and adaptive automation on human performance, situation awareness and workload in a dynamic control task. *Theoretical Issues in Ergonomic Science, 5*(2), 113-153.

Katidioti, I., Borst, J.P., van Vugt, M.K., & Taatgen, N.A. (2016). Interrupt me: External interruptions are less disruptive than self-interruptions. *Computers in Human Behavior, 63*, 906-915.

Luque-Casado, A., Perales, J.C., Cárdenas, D., & Sanabria, D. (2016). Heart rate variability and cognitive processing: the autonomic response to task demands. *Biological psychology, 113*, 83-90.

Miller, C.A., & Parasuraman, R. (2007). Designing for flexible interaction between humans and automation: Delegation interfaces for supervisory control. *Human factors*, *49*(1), 57-75.

Palinko, O., Kun, A.L., Shyrokov, A., & Heeman, P. (2010). Estimating cognitive load using remote eye tracking in a driving simulator. In *Proceedings of the 2010 symposium on eye-tracking research & applications* (pp. 141-144).

Parasuraman, R., Mouloua, M., & Molloy, R. (1996). Effects of adaptive task allocation on monitoring of automated systems. *Human factors, 38*(4), 665-679.

Parasuraman, R., Mouloua, M., Molloy, R., & Hilburn, B. (1993). Adaptive function allocation reduces performance cost of static automation. In *7th international symposium on aviation psychology* (pp. 37-42).

Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A model for types and levels of human interaction with automation. *IEEE Transactions on systems, man, and cybernetics-Part A: Systems and Humans, 30* (3), 286–297.

Parasuraman, R., & Wickens, C. D. (2008). Humans: Still vital after all these years of automation. *Human factors, 50* (3), 511–520.

Patterson, R. E. (2017). Intuitive cognition and models of human–automation interaction. *Human Factors, 59* (1), 101–115.

Prinzel III, L. J., Freeman, F. G., Scerbo, M. W., Mikulka, P. J., & Pope, A. T. (2003). Effects of a psychophysiological system for adaptive automation on performance, workload, and the event-related potential p300 component. *Human factors, 45* (4), 601–614.

Santiago-Espada, Y., Myer, R. R., Latorella, K. A., & Comstock Jr, J. R. (2011). The multi-attribute task battery II (MATB-II) software for human performance and workload research: A user's guide.

Sheridan, T. B., & Verplank, W. L. (1978). *Human and computer control of undersea teleoperators* (Tech. Rep.). Massachusetts Institute of Technology, Cambridge Man-Machine Systems Lab.

Stevens, C. A., Fisher, C. R., Morris, M. B., Myers, C., Spriggs, S., & Dukes, A. (2018, July). Cognitive Metrics Profiling: A Model-Driven Approach to Predicting and Classifying

Workload. In *International Conference on Applied Human Factors and Ergonomics* (pp. 236-245). Springer, Cham.

LIST OF ACRONYMS

ISR	Intelligence, Surveillance, and Reconnaissance
FMV	Full-Motion Video
AM	Autonomous Manager
SA	Situation Awareness
HMT	Human Machine Team
EEG	Electroencephalography