2008-08-07

On Autonomous Multi-agent Control in Wilderness Search and Rescue: A Mixed Initiative Approach

Benjamin C. Hardin
Brigham Young University - Provo

ON AUTONOMOUS MULTI-AGENT CONTROL IN WILDERNESS SEARCH AND RESCUE: A MIXED INITIATIVE APPROACH

by

Benjamin C. Hardin

A thesis submitted to the faculty of
Brigham Young University
in partial fulfillment of the requirements for the degree of

Master of Science

Department of Computer Science
Brigham Young University
December 2008

BRIGHAM YOUNG UNIVERSITY

GRADUATE COMMITTEE APPROVAL

of a thesis submitted by

Benjamin C. Hardin

This thesis has been read by each member of the following graduate committee and by majority vote has been found to be satisfactory.

Date

Michael A. Goodrich, Chair

Date

Dan Olsen

Date

Quinn Snell

BRIGHAM YOUNG UNIVERSITY

As chair of the candidate's graduate committee, I have read the thesis of Benjamin C. Hardin in its final form and have found that (1) its format, citations, and bibliographical style are consistent and acceptable and fulfill university and department style requirements; (2) its illustrative materials including figures, tables, and charts are in place; and (3) the final manuscript is satisfactory to the graduate committee and is ready for submission to the university library.

Date

Michael A. Goodrich
Chair, Graduate Committee

Accepted for the Department

Kent Seamons
Graduate Coordinator

Accepted for the College

Dr. Thomas W. Sederberg
Associate Dean, College of Physical and Mathematical Sciences

ABSTRACT

ON AUTONOMOUS MULTI-AGENT CONTROL IN WILDERNESS SEARCH AND RESCUE: A MIXED INITIATIVE APPROACH

Benjamin C. Hardin

Department of Computer Science

Master of Science

Searching for lost people in a Wilderness Search and Rescue (WiSAR) scenario is a task that can benefit from large numbers of agents, some of whom may be robotic. These agents may have differing levels of autonomy, determined by the set of tasks they are performing. In addition, the level of autonomy that results in the best performance may change due to varying workload or other factors. Allowing a supervisor and a searcher to jointly decide the correct level of autonomy for a given situation ("mixed initiative") results in better overall performance than giving an agent absolute control over their level of autonomy ("adaptive autonomy") or giving a supervisor absolute control over the agent's level of autonomy ("adjustable autonomy").

ACKNOWLEDGMENTS

Thanks to Dr. Michael A Goodrich for being a great advisor, Kayla for being a wonderfully encouraging wife, and inummerable friends, relatives, and lab partners who gave support.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Searching for lost people in wilderness terrain (Wilderness Search And Rescue, or WiSAR) is a large problem domain that is comprised of many smaller though related problems, such as building maps, managing logistics, and searching. Many of these smaller tasks have the potential to benefit from a team of multiple agents and, from a practical standpoint, many WiSAR tasks require a large number of agents. These agents can be a combination of human and robotic, but as the number of agents working on a task increases, so does an Incident Command Supervisor's (hereafter "supervisor") potential workload. The inflow of information increases, along with the attendant problems of managing, correlating, and applying the information. Coordinating and assigning tasks, managing the safety of teams, and strategizing becomes more complex and time-consuming. This can negatively affect overall performance as the supervisor has less situation- and change-awareness, more mental workload, and less time to spend managing the agents [20].

To address increased workload in scenarios such as this, a common solution is to give the agents autonomy. Autonomy can be characterized many different ways (discussed in detail later), but in general, *autonomy* is defined by what *tasks* the agents take or are assigned [32]. These tasks can be broadly defined, or broken down into as much detail as necessary. In a WiSAR scenario, the most basic high-level task is searching for a missing person. Subtasks may include dividing the search area into subareas, dividing the agents into teams, and assigning each team to a subarea. Sets of related or interconnected tasks are commonly referred to as *roles* [30]. An agent's *level of autonomy*, then, is the role or set of roles the agent fulfills. This leads to a key question: who or what determines the agent's level of autonomy?

If an agent is capable of at least some level of autonomous behavior, another question can be asked: can the agent's level of autonomy change? WiSAR scenarios are very fluid; challenges include complex terrain, agent availability, agent capability, human workload, weather, and many other contingencies. Each of these may be variable and can change throughout the duration of the situation; a search could move from flat to mountainous terrain, robotic agents could malfunction, or human searchers may only be available on a certain schedule. However, a human supervisor might not always be able to respond quickly or efficiently to changes in these variables due to workload or other demands on attention. One possible solution is to ignore the loss of performance and keep each agent at a static level of autonomy. If the agents have the ability to assume different levels of autonomy, however, one can consider two approaches: *adaptive autonomy* (allowing the agents to change their own level of autonomy) and *adjustable autonomy* (allowing a supervisor to explicitly change the agents' level of autonomy). In addition, a hybrid or "mixed initiative" solution can also be considered where the agent and supervisor collaborate to maintain the best perceived level of autonomy.

Assuming a set of WiSAR tasks that must be performed, and given a set of agents with the ability to perform those tasks autonomously, this thesis will compare the performance that results from the agents using adaptive autonomy, using adjustable autonomy, and using a mixed initiative combination. In the mixed initiative situation, there are many combinations of adjustable and adaptive autonomy to consider, so the focus will be limited to some "interesting" combinations, based on initial empirical experiments. In terms of a research hypothesis, this thesis will demonstrate that, given the constraints listed, performance in some metrics may decrease by using mixed initiative, but overall performance will improve by jointly taking advantage of an agent's capabilities and immediate but locally-detailed knowledge, and a supervisor's more complete but delayed knowledge.

1.1 Thesis Statement

Given a search task in a Wilderness Search and Rescue domain performed by a large heterogeneous group of agents, a mixed initiative system implementing both adaptive and adjustable autonomy performs better in a complex or high workload situation than a simply adjustable or adaptive system.

1.2 Thesis Organization

In the next chapter we review related literature. Chapter 3 contains an overview of the experiment and simulator design. Chapter 4 discusses results over several metrics. Chapter 5 outlines our conclusions and gives possible future research directions.

Chapter 2

Related Literature

Much has been written about search and rescue (SAR) [13] - [17], though less has been written about its sub-field of wilderness search and rescue (WiSAR) [40]. Rather than presenting a complete survey of all papers in these fields, this thesis will focus on related literature in more specific areas such as autonomy, teams, and coordination.

2.1 Autonomy

The definition of autonomy, as it applies to robotic actions, has been discussed in the literature in detail, with most definitions falling either into a function or task allocation category. Task allocation breaks tasks down into atomic sub-tasks [26]. Function allocation can be considered a weak form of task allocation [26], with basic functions being identified and being used no matter what the task breakdown is. One of the problems of function allocation is defining the basic functions, however, so we will use autonomy that employs task allocation. One definition of autonomy is that it is "an agent's active use of its capabilities to pursue some goal, without intervention by any other agent in the decision-making process used to determine how that goal should be pursued" [7]. Other definitions content themselves with simply equating autonomy to control [18], while others go into more detail of precisely what is being controlled [30].

The reason one would care about the definition of autonomy is that one might wish to apply it in concrete domains. Although autonomy can take an infinite number of forms and can be argued to be continuous [10], some people have attempted to cast it into practical roles. Endsley and Kaber describe four key roles: monitoring system status, generating strategies or options, selecting between those options, and implementing the chosen option [30]. In our WiSAR mission, checking to see if the UAVs are on path is an example of monitoring. Listing different teams that are able to move to that location and choosing one of them would be an example of generating and selecting, respectively. Performing the actual search would be an example of implementing. The *level of autonomy*, then, would be the combination of roles or superset of tasks that an agent takes.

Sheridan and Verplank first classified the different autonomous forms into discrete levels of autonomy [42], then later revisited it [38][41]. Endsley and Kaber also present a Level Of Autonomy (LOA) taxonomy. An overview of the development of various taxonomies can be found in [30]. More recently, there has been some work on defining levels of team autonomy [23], although this area is still new and this thesis will attempt to add to it.

Different levels of autonomy imply an ability of agents to exchange their current level of autonomy for a new one. Changing autonomy has variously been described as "mixed initiative" [21] - [22], "sliding autonomy" [11][25], "dynamic autonomy" [36], "adjustable autonomy" [10][21][22], or "adaptive autonomy" [5][6]. Some researchers make the distinction between which field the autonomy is being explored in; for example, what the aerospace and space robotics field often calls "adjustable autonomy," the artificial intelligence field frequently calls "mixed initiative" [21]. Other researchers consider the terms to mean the same thing.

Since the terminology is subjective, this paper will simply consider three terms: "adjustable autonomy," "adaptive autonomy," and "mixed initiative." This paper will define "adjustable autonomy" to mean a system of autonomy which requires a supervisor, monitoring system, or other outside influence to change the level. "Adaptive autonomy" will describe a system that allows the agent itself to change its own autonomy, perhaps based off various performance metrics or other trigger strategies. "Mixed initiative," then, is where both the supervisor and agent share in deciding the proper level of autonomy. These are meant to be high-level definitions. Precise implementation, role refinement, or task divisions may differ between domains.

2.2 Teams

Since this paper will be working with agents in a WiSAR domain, autonomy taxonomies of interest are those that can be scaled to large numbers of agents, or teams of agents. This in turn requires one to look at what exactly makes a team. Unfortunately, experimenting with large-scale teams can have a prohibitive cost, resulting in very little prior research about massively multi-robot situations. The majority of existing research deals with swarm robotics [8][33], which applies loosely at best to the search and rescue domain. Most other prior research in multi-agent teams addresses fewer than a dozen agents of two or three different types [1][31] with a large subgroup of research devoted to 11-agent robotic soccer teams [12][16]. Certain results found from these teams could potentially scale to a large WiSAR team, but other problems are unique to a large team.

In addition, much work has been done on the actual formation of teams or coalitions, including initial experiments in applying algorithms to real-life robots rather than software agents, in non-simulation scenarios [45]. However, this work focuses on peer-to-peer negotiating. While there may be some peer-to-peer negotiating inside a team while assigning low-level tasks in WiSAR, this paper will assume the supervisory method of team formation.

2.3 Metrics

Metrics for measuring human-robot interactions can be divided into three main categories: operator, robot, and system (measuring teamwork between human and robotic agents) [20]. Within these categories, there are several major metrics that are particular to each.

When discussing a hierarchical team organization, the idea of “span of control” [44] or “fanout” [15][29] is an important system metric. Because of time and attention constraints, a supervisor can effectively control a limited number of subordinates. This number may depend on the interconnectedness of the subordinates’ work and the level of performance that is desired or required, as well as other factors.

System metrics can be divided into quantitative and subjective measures. Quantitative measures include the efficiency and effectiveness of a team under various autonomous modes. Subjective measures can also be taken from all involved parties, in our case the team supervisor, including ease of control. In our experiments, we will also want to assess whether we appropriately utilized mixed-initiative.

Metrics specific to the robotic agents in our scenario include measurements of how aware the robotic agents are of other robotic or human agents, the humans working with them, and autonomy metrics such as neglect tolerance [37].

In addition, we consider two major measures of operator performance: workload and situation awareness [20]. Workload becomes important in a WiSAR domain because of its inherently hierarchical nature and life-critical tasks. A single supervisor may be responsible for overseeing many subordinates, with a workload inversely proportional to the level of autonomy granted or adopted by them. Situation awareness [30], on the other hand, can be negatively impacted by an increased level of autonomy of subordinates [3]. The question of whether increased autonomy results in an overall increase of operator performance will need to be addressed as well.

Chapter 3

Experiment Design

Two experiments were performed. The first experiment had a three by two design, comparing performance over three types of autonomy (adaptive, adjustable, and mixed initiative) and two levels of workload (high and low). A high level of workload was created by uniformly distributing 400 distracting items across the map, and a low level of workload was created by distributing 200 distracting items. Experiments were counterbalanced to avoid learning effects. 12 people voluntarily participated in the first experiment and performed four simulations each, resulting in 48 test cases. Subjects were compensated for their time. Data was collected between December 2007 and February 2008.

As we believed, the first experiment showed little difference in performance between high and low levels of autonomy. As a result, the second experiment adopted a uniformly high level of workload. The second experiment also had a three by two design, comparing performance over three types of autonomy (adaptive, adjustable, and mixed initiative) and whether the supervisor had expert knowledge (yes or no). To simulate expert knowledge, which our random pool of experimenters did not have, five "clouds" were placed on the map, corresponding to places where expert knowledge would indicate had the highest probability of containing missing person (Figure 3.2). These clouds did not equate to perfect knowledge, however: one of the five clouds was not paired with one of the five missing people. Results were compared between scenarios where the experiment had the expert knowledge clouds, and scenarios where they did not. 18 people participated in the second experiment, resulting in 72 test cases. Subjects were compensated for their time. Data was collected between February and March 2008.

At the beginning of each experiment, subjects participated in a training mission, lasting approximately 10 minutes, which introduced them to the functions and controls of the simulator, as well as explaining the purpose of their mission.

3.1 Arenas

One issue when considering or comparing results is reproducibility. Both the original experimenter and other experimenters need to be able to reproduce the results for true validation. Whether live or in simulation, however, there are many factors that can affect performance, making it difficult to determine which factors influenced the results the most. One significant factor can be the environment in which the experiments are run, meaning the particular structure of the wilderness area (often called a field in this thesis) in which the search takes place. With an eye to facilitating reproducibility of wilderness search and rescue simulations, we have created several test fields. An example is shown in Figure 3.1.

A similar thing was done in urban search and rescue [27]. Three standardized test arenas simulate various stages of urban environment complexity. These arenas allow disparate teams to test the performance of their robotics teams while still making it acceptable to compare results.

While these urban arenas can be physically built to allow live experimentation, a wilderness arena brings additional difficulties. First, given the vast sizes of wilderness areas, it is impossible to reproduce a wilderness arena at will. Second, using a standardized live wilderness setting can lead to training overfit. However, since our initial wilderness search and rescue experiments are performed in simulation, we can at least create a simulated wilderness arena that will allow us to reproduce our results and compare performance with other simulations that run on the same arena. Although our experiments were run on four unique maps held constant for the purposes of comparison between experiments, a more general technique would be to randomly generate maps based on percentages of passable

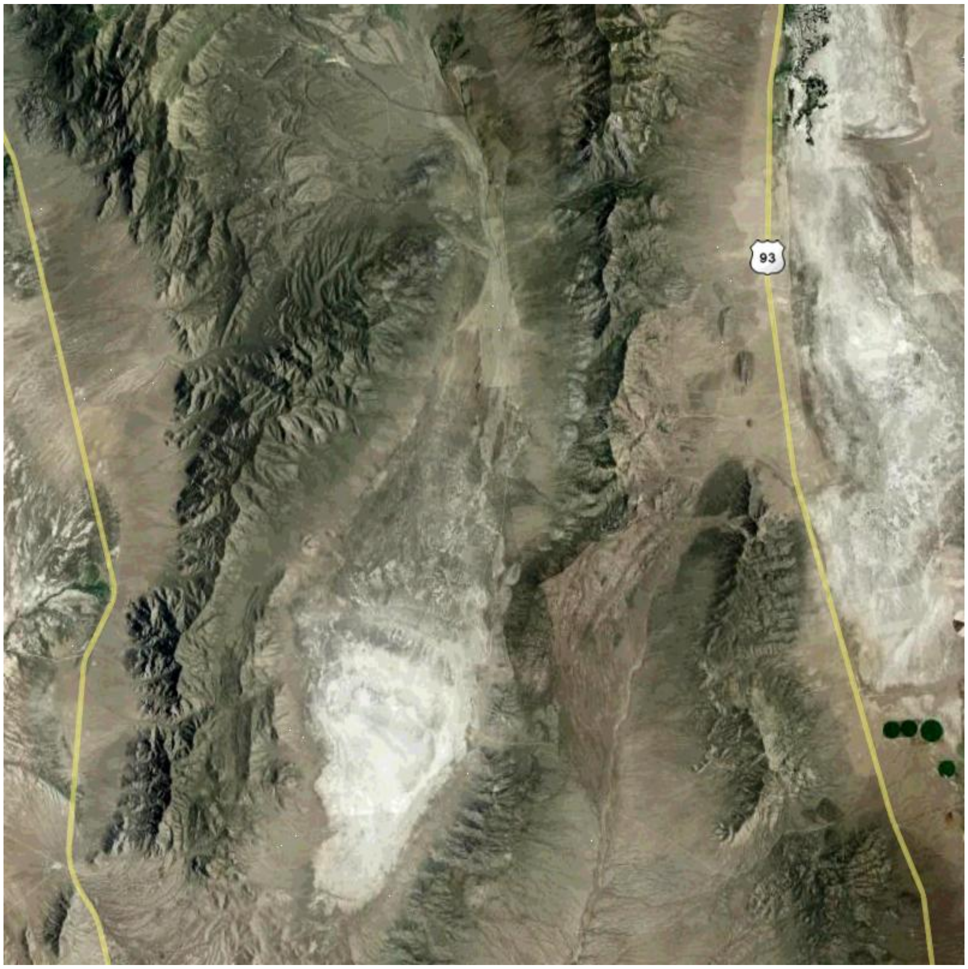


Figure 3.1: Example Test Field

or impassible terrain, random placement of obstacles, and random placement of missing people to avoid overfitting to a particular situation.

When creating the urban test arenas, previous research considered several categories of agent capabilities: mobility, sensing, knowledge representation, planning, autonomy, and collaboration. Once these components were identified, arenas were created to tax the agents employing them. We can take ideas from these urban arenas, while adding necessary new factors to preserve ecological validity for a wilderness arena. We will look at each component in turn.

- Mobility. Travel through an urban environment will be shaped by rubble, elevation changes through stairs and ramps, stability of the travel surface, size and shape of passageways, and even the material of the travel surfaces. In wilderness search and rescue, travel may be hindered by obstacles such as cliffs, water, heavy brush, or elevation changes; or could be helped by trails or watercourses. However, the urban arenas presuppose ground robots while a wilderness search might include ground and air searchers, with specialty teams for water and cliffs or other technical terrain.

- Sensing. In urban search and rescue, robots may have to sense acoustic, thermal and visual signals to locate victims. These signals may be distorted or noisy due to the environment. The same holds true in wilderness search and rescue, although signals may need to be sensed at greater distances. In addition, dog teams or robotic agents that can sense chemical markers may be able to access scent signals.

- Knowledge representation. In both urban and wilderness search and rescue, a robot may have to make maps and mark locations of obstacles, missing people, unsearched areas, or exits. Similarly, a wilderness arena should be complex enough to require map-making, albeit on a larger scale.

- Planning. Partly due to the larger spatial areas that must be covered and the additional issue of victims possibly changing their spatial location, planning is just as vital in wilderness search and rescue as in urban. Chokepoints should be inspected and possibly monitored; constraints can be applied by patrolling borders.

- Autonomy. Communication may be impossible, noisy, or inconsistent due to terrain, faulty equipment, or other reasons. A supervisor may not have enough time to give constant directions. Situations may arise that require immediate reaction from the searcher to avert danger or facilitate rescue. In these cases, the agent may have to employ autonomy; whether adaptive, adjustable, or mixed-initiative.

- Collaboration. In many wilderness search and rescue scenarios, search is less a matter of carefully negotiating a technical and dangerous situation and more a matter of searching a large area as quickly as possible. Regardless, a wilderness search and rescue can benefit from a large number of searchers. In this case, peer-to-peer collaboration between agents or collaboration between a supervisor and agents is essential.

Several things set wilderness search and rescue apart from urban search and rescue. In an urban disaster scenario, there are often multiple victims, their location is known to within a few meters, and their location does not change significantly over the course of the rescue or recovery. In a wilderness scenario, there is often just a single victim. The victim is possibly moving, and their travel decisions may not be based on logic. There often are not enough searchers to effectively cover the search area. Rather than a single all-purpose searcher commonly used in urban scenarios, there may be need for specialized teams to handle different terrains or situations: aquatic teams of rivers or lakes, climbers or rapellers for cliffs, and so forth. These factors all combine to make it difficult to statistically measure and average the performance of a search system. As a result, a wilderness simulator may be forced to sacrifice some ecological validity for statistical significance by implementing multiple victims. In our case, we scattered five identical missing people across the map, each approximately equidistant from the starting location of the searchers. This removed statistical anomalies arising from a fortunate or unfortunate choice of initial direction by the search supervisor, allowing us to average the amount of time to find each person.

In addition, we gave each of the five missing people 16 identical items (Figure 3.5). Each of the people, in the course of their simulated wandering, lost a subset of 10 of those 16 items, scattered following a Gaussian distribution around the final location of the person. Those items, when found (Figure 3.6), allowed the search supervisor to focus their attention, knowing the missing people were in the vicinity. Finally, "distracting" items were scattered across the map following a uniform distribution, providing a secondary task and additional workload for the search supervisors as they classified the items as "good" items or "distracting" items.

The terrain we chose for the four scenarios ranged from primarily flat to extremely rugged. Mobility of the search agents was affected by impassible terrain of varying degrees of complexity. For our experiments, we chose to preserve ecological validity by implementing only basic sensing, giving our simulated searchers simple proximity-based sensors. In addition, sensing was deliberately made imperfect. For our first set of experiments, search grids of varying density were implemented, with probabilities of discovery based on previously gathered data [40].

For a search grid with 30 meter spacing between searchers, probability of detection was 50 percent. Spacing searchers 18 meters apart resulted in a 70 percent chance, and spacing them 6 meters apart resulted in a 90 percent chance of detection. Search supervisors were required to decide on a trade-off between a fast search that covered a large amount of ground and took fewer searchers but resulted in a lower probability of detection; or a detailed search that took more searchers, covered less ground, but had a higher chance of finding a missing person. Because of the probabilistic nature of item or missing person discovery, supervisors could potentially have to make several passes over the same search area to recover all items or find the missing person.

The peoples' missing items, distracting items, and the people themselves all had the same probability of discovery. Later experiments cared less about measuring the trade-off between the different search patterns, and we removed the option of the 18 meter search pattern while retaining the 30 meter and 6 meter patterns.

Each experiment gave the supervisor control of 200 searchers. The searchers were giving the capability to act autonomously, and finding the method of choosing their level of autonomy that resulted in the best performance was our primary objective. The three methods we looked at were adaptive, adjustable, and mixed initiative.

3.2 Autonomy

In adaptive scenarios, the searchers chose their own level of autonomy, with no direct input from the supervisor. They would begin with a general 30 meter sweep, focusing on the center of the search area before fanning out to the more distant locations. In our experiments we called this the "high" level of autonomy, since they had no direct input from the supervisor. The supervisor's task was limited to classifying items the searchers found; although by choosing to "keep" items they could still indirectly influence the searchers' level of autonomy.

The moment the supervisor elected to retain an item by clicking the "keep" button, a subset of searchers would create a smaller, more detailed 6 meter search area centered around the item and adaptively change their level of autonomy to a "medium" level to conduct a detailed search. At this medium level of autonomy, they would continue to search the search area until they found the missing person and the supervisor classified it as such. After they discovered the missing person in the detailed search area and the supervisor classified the person as such, the searchers would immediately return to high autonomy and continue a broad search for the remaining missing people. In the scenarios that employed adaptive autonomy, the searchers never chose to go to low autonomy.

In adjustable scenarios, the search supervisor had to initiate all changes in the searchers' level of autonomy. A menu was available which allowed the supervisor to grant high autonomy. This would give the searchers permission to create their own search areas and then search them, successively moving on to new search areas as they completed the previous one.

Whether the searchers were at high or low autonomy, if the supervisor created a "small" (defined as less than 22,500 square cells, e.g. 150 by 150 cells if square) search area around a discovered and classified item, then assigned searchers to it, the searchers would immediately go to a medium level of autonomy. At this medium level, they would continue searching until they discovered the missing person and the supervisor classified it as such, repeating the search as many times as necessary. After the missing person was located and classified, the searchers would complete their current search area to discover as many backpack items as possible, then return to a low level of autonomy and await further instructions. When they had low autonomy, the searchers would perform a search of an assigned area, then await further instructions.

In mixed initiative scenarios, the searchers would start with a high level of autonomy, and return to high autonomy whenever possible. If the search supervisor assigned them to a search area, they would immediately assume a low level of autonomy and search the area. When the search was complete, the searchers would return to high autonomy and choose their own subsequent search area. Similarly, as soon as they completed a detailed search area at medium autonomy, they would also return to high autonomy, exactly like agents operating under adaptive autonomy.

A supervisor could manually create a small (arbitrarily defined) search area around a discovered and classified item. Assigning searchers to it would result in them assuming a medium level of autonomy, but as soon as their search was complete, they would immediately return to high autonomy. However, while operating under mixed initiative, if an item was discovered and classified as a backpack item, the searchers would also take the initiative and create their own small search area around them, moving to medium autonomy for the duration of the search of the small area. Identical to adaptive autonomy, they would search the area at medium independence until the missing person was discovered and classified, then return to high autonomy.

3.3 Metrics

Prior to and during the experiments, several key metrics were identified. These fell approximately into three categories: primary task performance, secondary task performance, and management overhead/workload. We will discuss different metrics for each of these categories in the subsections that follow. Note that we assigned a level of importance to each metric, although this assignment was very subjective. Depending on the exact goals of a given task or scenario, the relative importance of each metric may be adjusted.

3.3.1 Primary Task Performance

Task performance metrics focus primarily on the results. Most of these metrics revolve around the primary task of finding and classifying the missing people, and the secondary task of finding as many backpack items as possible. In addition, supporting or related tasks are also considered, such as covering as much ground as possible. We will look at the key primary task performance metrics first.

- Average number of missing people classified. (High Importance) This is one of the more significant metrics. In the end, the main measure of success in a wilderness search and rescue domain is whether the missing person is recovered. Finding as many of the five missing people as possible was the supervisor's primary task.

- Probability of success. (High Importance) This is a metric that is used in the Search and Rescue community, and relates the "probability of area" (the probability that the missing person is contained in a given search area) with the "probability of discovery" (the probability of actually seeing the missing person when searching near them). As this metric is more complex than others, it will be described in more detail later.

- Percentage of times all five missing people were found. (Medium Importance) Although the experiment was deliberately designed to make discovery of all five lost people difficult, it did still happen. This is another strong metric.

- Average time to find all five missing people. (Medium Importance) This metric is valid only for scenarios where all five missing people were found.

3.3.2 Secondary Task Performance

Secondary tasks are any tasks that do not directly involve finding the missing people. The most important secondary task is finding as many of the backpack items as possible, but additional ones are also outlined below.

- Average number of backpack items classified. (Medium Importance) Experimenters were told that their primary task was to find all five of the missing people, but they had a secondary task of finding as many of the 50 (10 per missing person) backpack items as possible. This number can be broken into two sets of scenarios: those where the searcher first accomplished their primary task of finding the missing people, and those where the searcher performed both primary and secondary tasks in tandem.

- Average number of distractors classified. (Low Importance) This value should be proportional to the amount of terrain covered.

- Average searcher distance, or number of cells moved. (Low Importance) For the purposes of our experiments, each cell was considered to be six by six meters. This allowed us to easily translate the 30-, 18-, and 6-meter search spacing into 5-, 3-, and 1-cell spacing. However, by doing this, we must consider "time" in the experiment to be highly sped up, since in our simulations, the travel speed was approximately 4 cells per second. Additionally, it must be noted that the traveling speed of the agents was not guaranteed to be ecologically valid given the terrain they were on, i.e. the searchers' speed remained constant regardless of terrain. Future work may wish to employ a more precise simulation of time and space. Similar to the next metric (average timesteps inactive), both adaptive autonomy and mixed initiative should result in approximately equal values for the average searcher distance, since the agents have the initiative in both cases to continually search without being required to stop moving to wait for instructions.

- Simple coverage. (Low Importance) This metric gives the percentage of the searchable terrain that was jointly covered by the searchers. This number is not weighted by the most probable locations of the missing people nor the depth to which the terrain was searched, so it may be of limited usefulness.

3.3.3 Management Overhead/Workload

Management overhead and workload metrics measure demands on or performance of the supervisor. Several of these metrics are related to task performance metrics, and may even involve the primary or secondary tasks, but their emphasis is on the abilities and responses of the supervisor rather than the task itself. Primary workload, as we measured it, included creating search areas, selecting agents, assigning agents a destination on the map, and assigning agents to search areas.

- Average timesteps inactive. (Low Importance) This value only applies to adjustable autonomy, since the agents operating under adaptive autonomy or mixed initiative have the ability to instantly choose a new task when they complete their previous one, rather than going inactive. Under adjustable autonomy, however, agents may be idle while waiting instruction from the supervisor.

- Average number of timesteps taken to classify items. (Medium Importance) Time is counted from when an item is discovered by the searchers and placed in the classification queue, to the point when the supervisor chooses to keep or reject the item. This metric is a factor of the amount of workload that results from tasks other than classifying items, such as creating search areas and directing the searchers. Adaptive autonomy should result in the best value for this metric, since the supervisor's only task is to classify items. Mixed initiative and adjustable autonomy should both be approximately equal.

- Workload. (Medium Importance) Workload can include creating search areas, selecting searchers, assigning searchers to search areas, changing searchers' autonomy level, and classifying items.

- Average time between finding the first item belonging to a particular missing person and finding the missing person. (Medium Importance) This is a factor of both the response time of the controlling entity (the supervisor in an adjustable autonomy situation, the search agents themselves using adaptive autonomy, or both while using mixed initiative) and their strategy (the number of agents assigned to the search area, etc.).

- Percentage of items correctly classified. (Low Importance) This value does not have much significance, since all scenarios should have a similar, almost perfect score. All of the items are available to view so there is no memorization necessary. After classifying items as "good" or "bad," there was no reversing the decision. There was no indication during the experiment whether they classified items correctly or incorrectly.

3.3.4 Probability of Success

One important primary task performance metric is the *probability of success*.

Probability of success can be difficult to interpret due to the lack of a formal technique for either discretizing a search area or weighting each discretization. In its simplest form, it is calculated by multiplying the probability of detecting the missing person ("probability of detection") given a certain search pattern by the probability that the missing person was in the search area to begin with ("probability of area"). Calculating the probability of success becomes more difficult when several search techniques with different probabilities of detection are used, or a search area is divided into sub-areas with different probabilities of detection due to different terrain or other factors. In addition, the sub-areas can have different probabilities of containing the missing person, which probabilities can be drawn from expert knowledge, historical or statistical data or knowledge of the missing person's habits, capabilities, or personality.

When the search area is discretized, the probability of success for a given area is found by summing the product of each discretization's probability of detection and probability of area, then dividing by the maximum possible probability of detection multiplied by the probability of area. In our experiments we used a linear decay distribution around each missing person to weight our probability of area, ranging from a probability value of "1" directly over the missing person to a probability value of "0" at a radius of 50 cells from the missing person.

3.4 Simulator

The primary view of the main window (Figure 3.2) shows the location of the 200 searchers (the small yellow squares), the search areas (the large box with the diagonal lines), and a satellite view of the terrain. The five black clouds represent expert knowledge of the supervisor, and indicated areas that the supervisor's experience and intuition mark as the locations with the highest probability of containing the missing person.

Figure 3.3 shows an accessibility or simplified terrain map. Brown regions are mountainous and considered inaccessible; blue regions indicate water and are likewise inaccessible. Grey lines indicated roads or trails.

Figure 3.4 shows the action window. This allows the supervisor to choose between the two main actions: selecting and moving searchers, or creating search areas. After choosing the action marked "Create Search Area," the supervisor may use the mouse to draw one or more search areas of any size on the map. When the action titled "Select and Move Searchers" is chosen, a supervisor may use the mouse to highlight a number of searchers. Clicking on a location on the map or on a search area will command the searchers to move to that location or search the given search area, respectively.

The two green columns marked "Searcher Activity" and "Item Count" give the supervisor a quick, high-level look at their current performance. As the searchers become

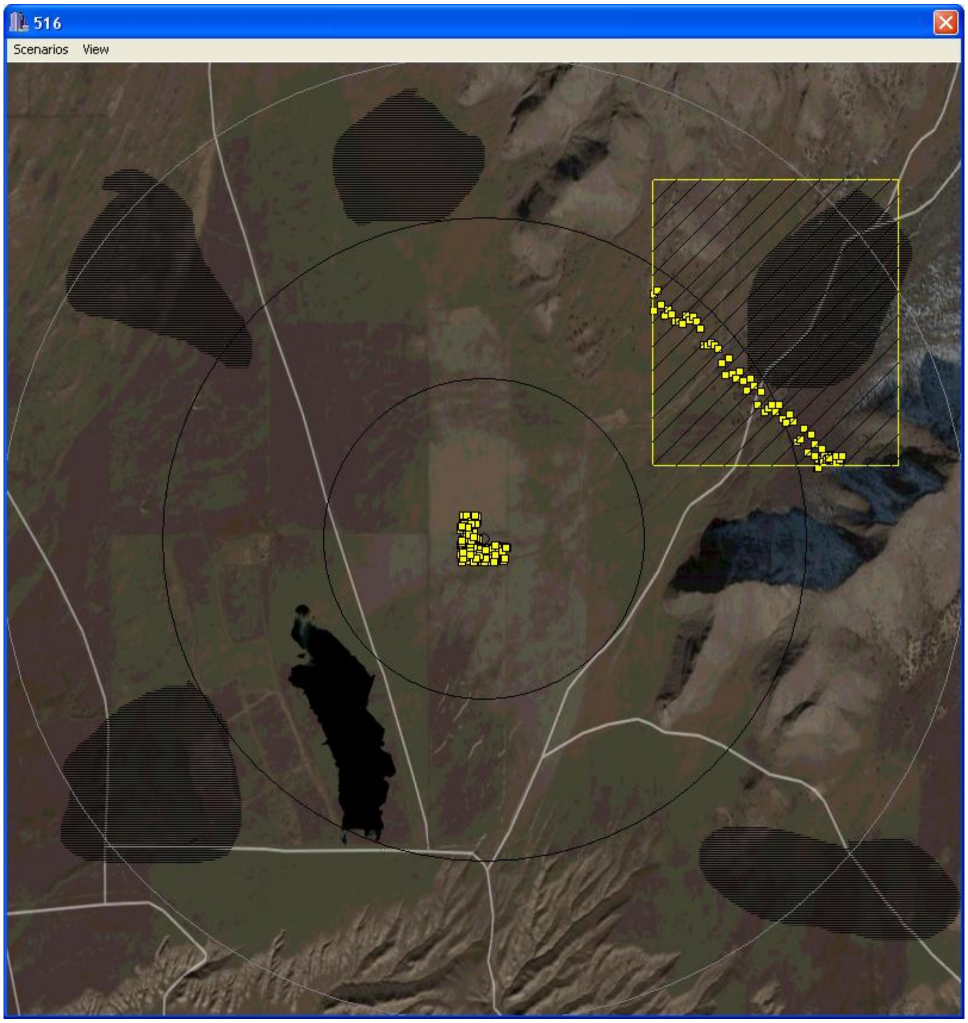


Figure 3.2: Simulator Main Window with Expert Clouds

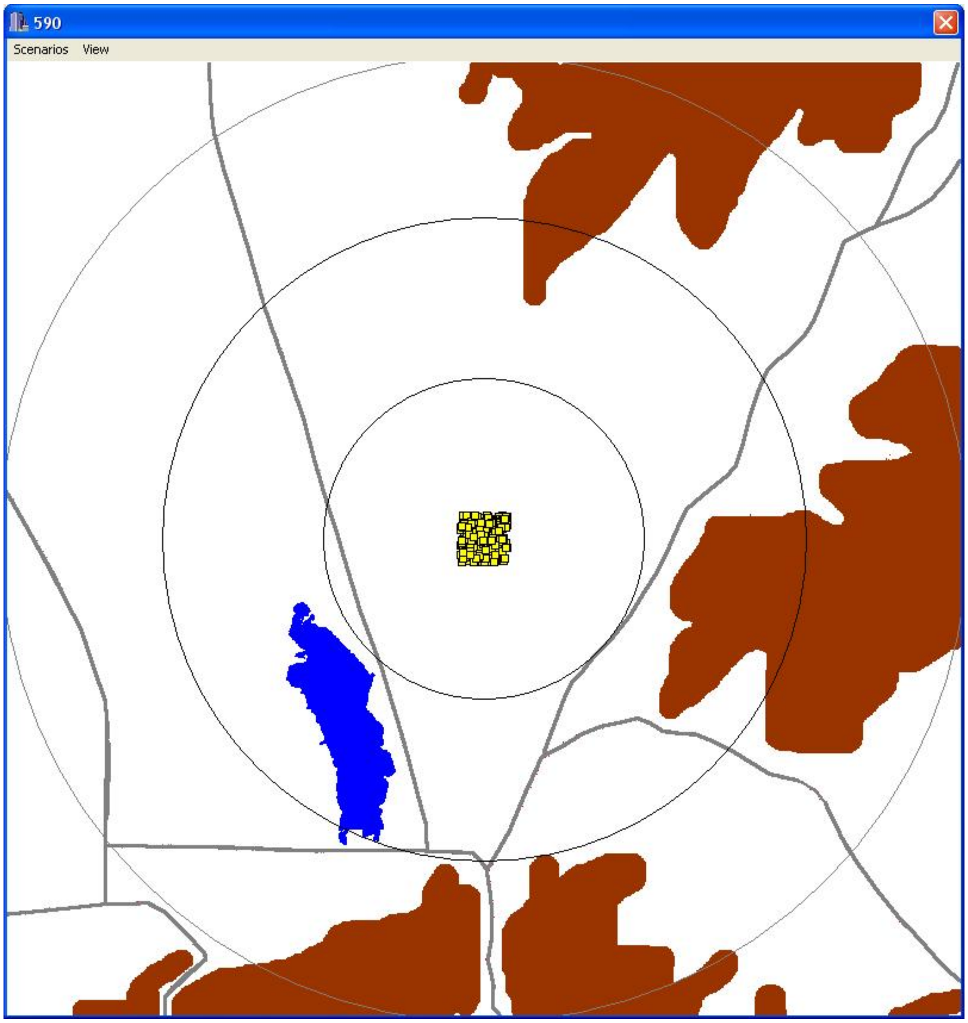
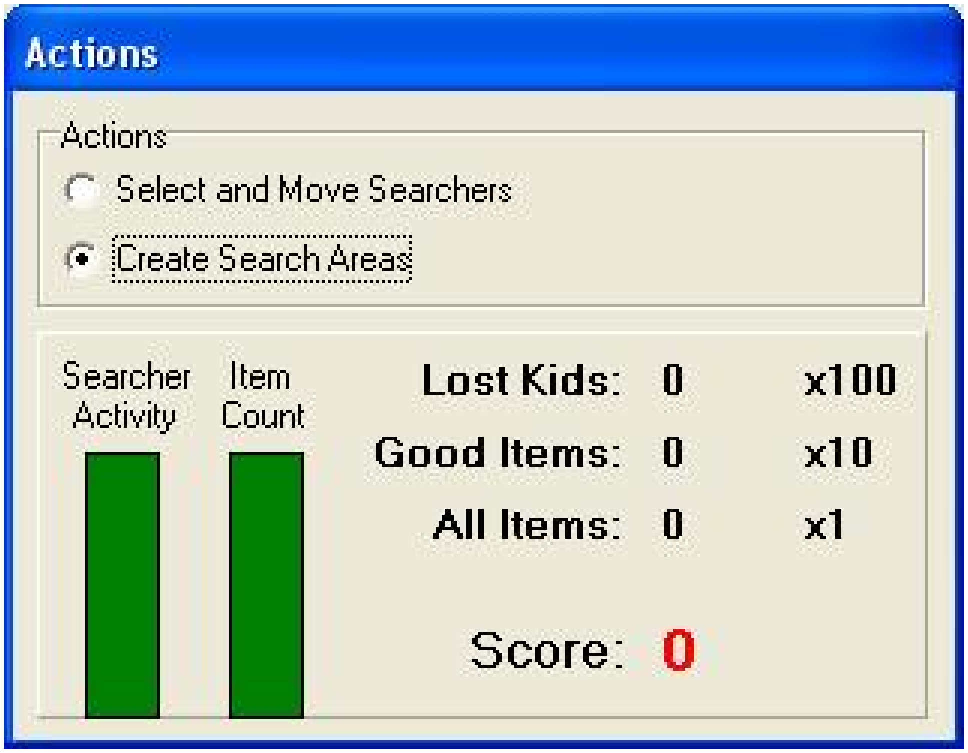


Figure 3.3: Simulator Main Window, Terrain View



Actions

○ Select and Move Searchers

● Create Search Areas

Lost Kids:	0	x100
Good Items:	0	x10
All Items:	0	x1

Score: 0

Figure 3.4: Action Window, Activity Monitor, and Scorecard

idle, the column marked "Searcher Activity" will drop, turning yellow to indicate that a significant number of searchers are inactive, and finally red to indicate that the majority of the searchers are idle. Similarly, the column marked "Item Count" will turn yellow as found items begin accumulating in the found items window (discussed below), and finally red when there are a large number of items that are waiting classification.

The "score" was merely used to allow participants to compare themselves against other participants, encouraging them to perform at the best of their abilities. Each discovered person awarded the subject with 100 points, each backpack item earned them 10 points, and each false item correctly classified as such earned them one point. This had the added benefit of continually reminding them of their priorities and the approximate weight between their primary task of locating the missing people and their secondary task of discovering as many backpack items as possible. This score was not used for anything other than subject encouragement.

At the top of the backpack window (Figure 3.5) was a picture of the missing person. All five of the missing people were considered "identical quintuplets." Below this picture were

16 additional pictures of the contents of the missing person's backpack. All five people had identical backpack items. Each person lost 10 of their 16 items, chosen at random.

Items that the searchers discovered as they traveled or searched an area were put in the found items window (Figure 3.6). Each item had two buttons, marked "Keep" and "Reject." Participants were instructed to keep items that came from the peoples' backpacks, and reject distracting items. Backpack items that were discovered looked identical to their pictures in the backpack window: no interpretation was necessary. As backpack items were correctly kept or distracting items were incorrectly kept, a small white marker was placed on the map to indicate the location the item was discovered at. No indication of false classification was given. When a person was discovered and subsequently classified as a missing person, a green marker was placed on the map to indicate his location.

Figure 3.7 shows a map with the locations of all the missing people and backpack items indicated, identical to how they would appear if the supervisor had correctly located all of them. The backpack items were distributed following a Gaussian distribution. The missing people are indicated by small green circles (emphasized by red arrows in this image), while the backpack items are indicated by small white circles. The initial location of the searchers is also indicated.

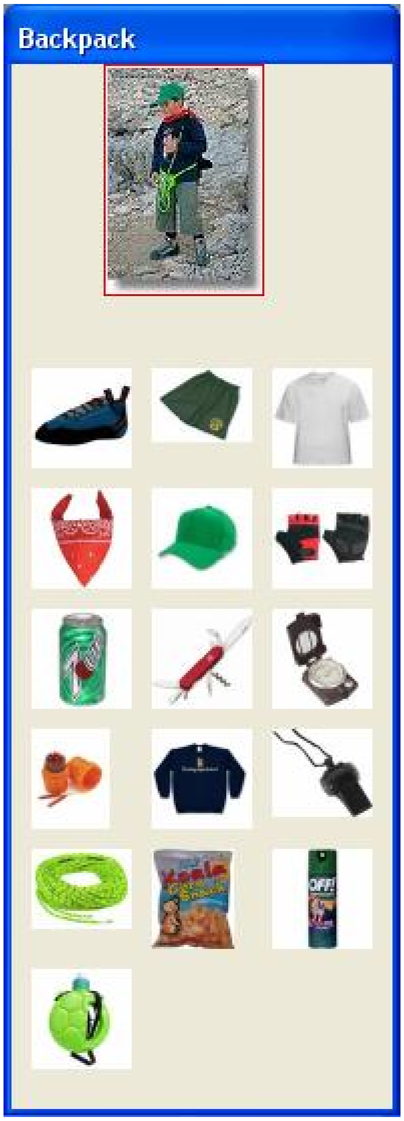


Figure 3.5: Simulator Backpack Window

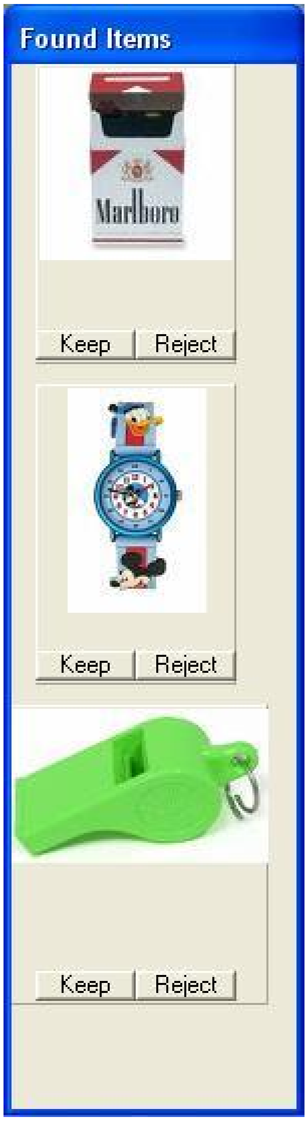


Figure 3.6: Simulator Found Items Window

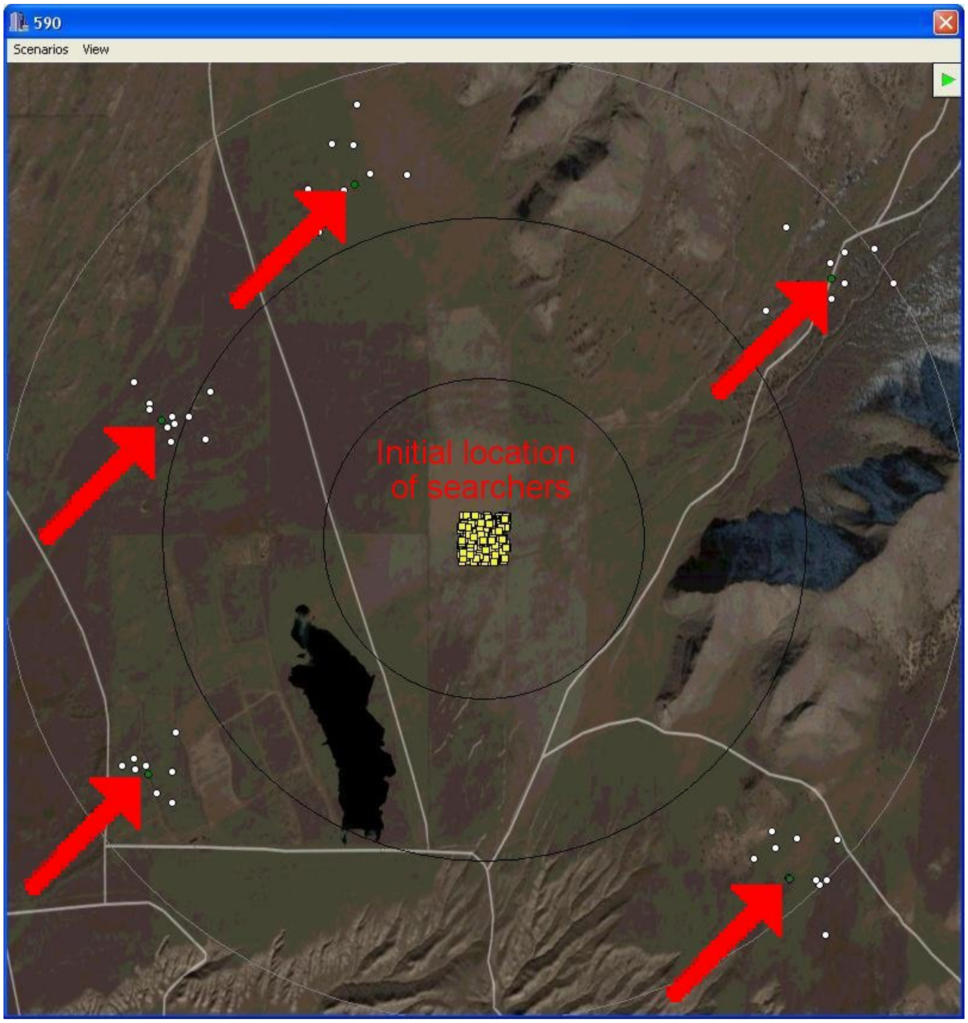


Figure 3.7: Main Window with Targets and Backpack Items Indicated

Chapter 4

Results

Two formal experiments were performed. Experiment one was a preliminiary experiment designed to help choose the level of workload for the main experiment. Following that, additional data was gathered from several users to focus and validate certain design changes, then the final, more informative experiment was performed.

4.1 Experiment One: High vs. Low Workload

The first experiment was designed to show whether there was a statistically significant difference in performance when supervisors were given a high workload versus a low workload. In this case, workload was defined as the number of distracting items distributed on the map. Denser distributions led to more distracting items being discovered, forcing the supervisor to perform more classifications. A high workload was defined as 400 distracting items scattered across the map following a uniform distribution, while a low workload was 200 distracting items.

Whenever it was established (using a two-way analysis of variance) that there was a statistically significant difference across one of the independent variables (workload or autonomy), we looked at the significance of certain pairwise combinations of autonomy type/workload using t-tests. While looking at differences across workload, we looked at combinations that differed across workload. Similarly, while looking at differences across autonomy, we looked at combinations that differed across autonomy.

The three significant comparisons while looking at differences across workload were adaptive autonomy/high workload versus adaptive autonomy/low workload, adjustable autonomy/high workload versus adjustable autonomy/low workload, and mixed initiative/high workload versus mixed initiative/low workload. In subsection 4.1.1 these will be referred to as the “three considered comparisons.”

The six significant comparisons while looking at differences across autonomy were adaptive autonomy/high workload versus adjustable autonomy/high workload, adaptive autonomy/high workload versus mixed initiative/high workload, adjustable autonomy/high workload versus mixed initiative/high workload, adaptive autonomy/low workload versus adjustable autonomy/low workload, adaptive autonomy/low workload versus mixed initiative/low workload, and adjustable autonomy/low workload versus mixed initiative/low workload. In subsection 4.1.2 these particular six comparisons will be referred to as the “six considered comparisons.” All F-values and p-values (significance values) from the analysis of variance can be found in Appendix C, while t-values can be found in Appendix A.

4.1.1 Differences Across Workload

Of the metrics considered, only one resulted in a significant difference across workload (F-value of 5.04, p-value of 0.032). While using adjustable autonomy, the average length of time to classify items was significantly higher when there was a high level of workload. Using mixed initiative, the difference in performance between high and low workload was even greater (Figure 4.1). Looking at the three considered comparisons individually, t-values indicate that each is statistically significant.

4.1.2 Differences Across Autonomy

Several metrics had a statistically significant difference across autonomy. Average searcher distance, simple coverage, and average number of items classified all differed across au-

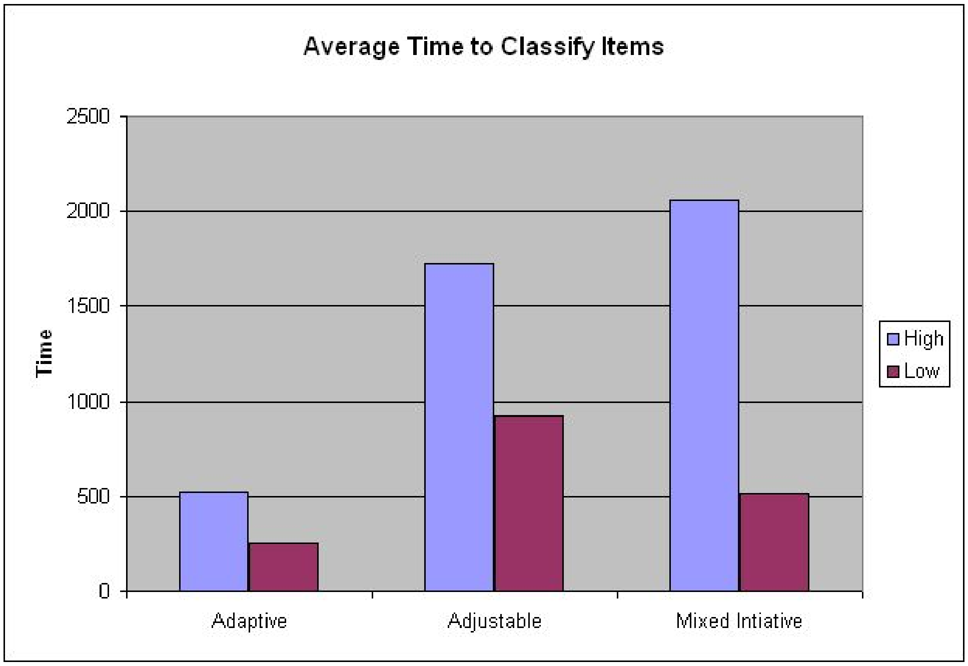


Figure 4.1: Average time to classify items (high vs. low workload)

tonomy. Incidentally, these are the three metrics classified as “low importance” in Table 5.1.

Average searcher distance had an F-value of 13.29 and p-value of 0.0001. T-values indicate that the differences between all six considered comparisons were statistically significant except the comparison between mixed initiative/high workload and adaptive autonomy/high workload. From Figure 4.2, however, we can see that even though the difference between mixed initiative/low workload and adaptive autonomy/low workload was statistically significant, it still was not great compared to the difference between adjustable autonomy and either adaptive autonomy or mixed initiative. When searchers using adjustable autonomy completed a task, they became idle until given a new one while searchers using adaptive autonomy or mixed initiative were able to pick a new task and continue moving. This resulted in searchers using adaptive autonomy and mixed initiative covering significantly more ground. Note that this says nothing about whether they were searching in locations that had a high probability of containing the missing people.

The difference in performance in simple coverage between searchers using adaptive autonomy and searchers using adjustable autonomy was proportional to the difference in

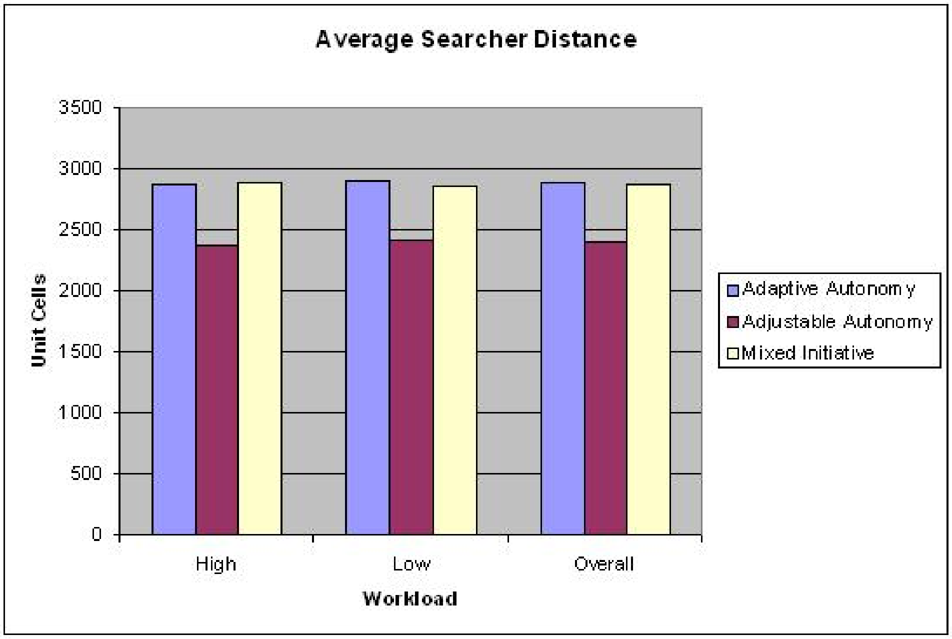


Figure 4.2: Average searcher distance (high vs. low workload)

average searcher distance. Searchers using mixed initiative performed worse than adaptive autonomy, but better than adjustable autonomy. The key to the difference is redundancy; while searchers using adaptive autonomy covered a lot of ground, searchers directed by a supervisor tended to focus on certain areas (presumed by the supervisor to have a high-probability of containing the missing people). This resulted in less total ground covered, but a deeper coverage of the searched areas. The difference across autonomy in this metric was significant over all six considered comparisons, and had an F-value of 6.5 and p-value 0.0044.

The final significant metric we looked at was the average percentage of the total number of items classified (Figure 4.3), with an F-value of 5.6 and a p-value of 0.0084. It was significant over all six considered comparisons except the comparison between adjustable autonomy/high workload and mixed initiative/high workload. It shows that adaptive autonomy resulted in the largest percentage of items classified; adaptive autonomy's larger area of coverage resulted in a larger percentage of the distracting items being found. The difference between high and low workload was not statistically significant (t-value of 0.03 and p-value of 0.05). At low workload, however, it was statistically significant that

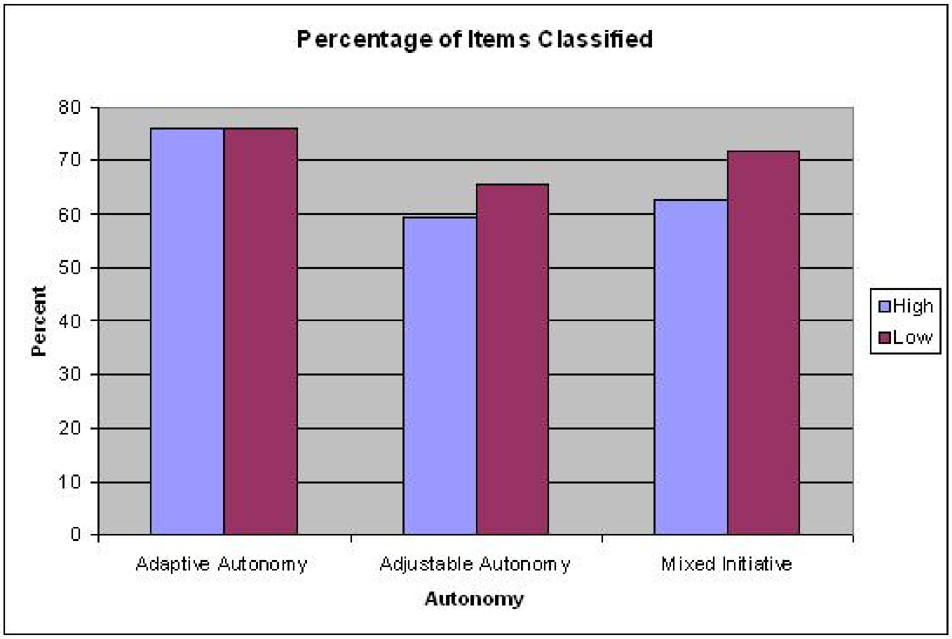


Figure 4.3: Percentage of items classified (high vs. low workload)

supervisors using mixed initiative were able to classify more items than supervisors using adjustable autonomy. As previously mentioned, the improvement of mixed initiative over adjustable autonomy at high workload was not statistically significant.

4.1.3 Two-Way Interactions

There were no significant two-way interactions between workload and autonomy in the metrics considered.

4.1.4 Discussion

Overall, the level of workload did not significantly affect performance. The amount of time it took to classify items was dependent on the level of workload, but this metric is not of high importance (Table 5.1). Also, the average searcher distance, coverage, and number of items classified were dependent on type of autonomy, but these three metrics are of small importance (Table 5.1). From this we can conclude that the level of workload and autonomy type is of minimal importance as far as performance is concerned.

As a result, we chose to perform all future experiments at a high level of workload.

After this initial study was completed, we experimented with making the search more difficult. This was done by decreasing the amount of ground searched for each movement, as well as decreasing the probability of discovery for each search that was performed. This second experiment will be looked at in more depth.

4.2 Experiment Two: Expert vs. Non-Expert Supervisor

Each of the three types of autonomy provided a unique set of capabilities and method of task execution, and as a result we would expect each to perform well in different metrics. We will look at the categories of primary task performance, secondary task performance, and management overhead/workload in turn.

Two-way analysis of variances run on each metric indicated that (a) none of the metrics showed statistically significant differences between expert and non-expert supervisors, (b) all the metrics showed a statistically significant difference across autonomy types, and (c) none showed any statistically significant two-way interactions. As a result, we can infer that the expertness of the supervisor did not affect the measured performance, and the type of autonomy did. F- and p-values from the analysis of variance tests can be found in Appendix D, and data and t-values can be found in Appendix B.

There were three exceptions; simple coverage was statistically significant across expert level, autonomy type, and had two-way interactions, while the average time to find at least one backpack item from both four and five missing people was statistically significant across expert level and autonomy type. These will be discussed in-depth later.

4.2.1 Primary Task Performance

The searchers' primary task was to locate the missing people. Speed, efficiency, and completeness in their search facilited fulfilling this task, as did searching in areas that had a high probability of containing the missing people. We will look at several metrics that

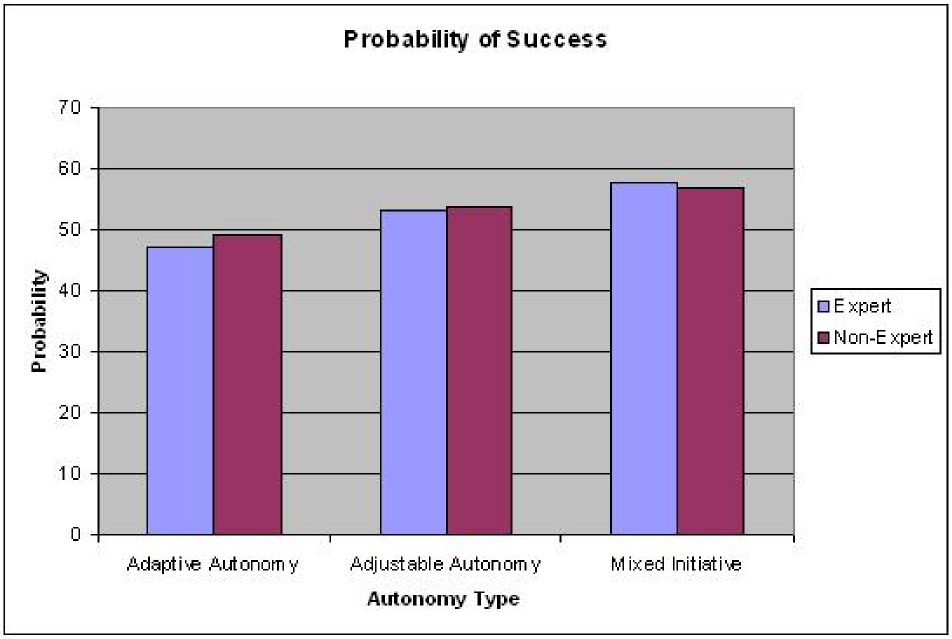


Figure 4.4: Probability of Success

relate to performance in the primary task.

Probability of Success

Probability of success (described in Section 3.3.4) was one of the most important high-level metrics to consider. Supervisors controlling searchers who employed mixed initiative had the highest probability of success at 57.5 percent for expert supervisors and 56.9 percent for non-expert searchers, although an analysis of variance showed that there was no statistical significance to the difference between expert and non-expert searchers over any autonomy type (F-vale of 1.13, p-value of 0.2919). Adjustable autonomy had the second highest probability of success, while adaptive autonomy came in last (Figure 4.4). This difference across autonomy type was statistically significant (F-value of 89.53, p-value of 0.0001).

Missing People

After looking at the high-level metric of probability of success, we can look at primary task performance metrics that directly involve finding the missing people (our primary goal) or finding their backpack items (our secondary goal).

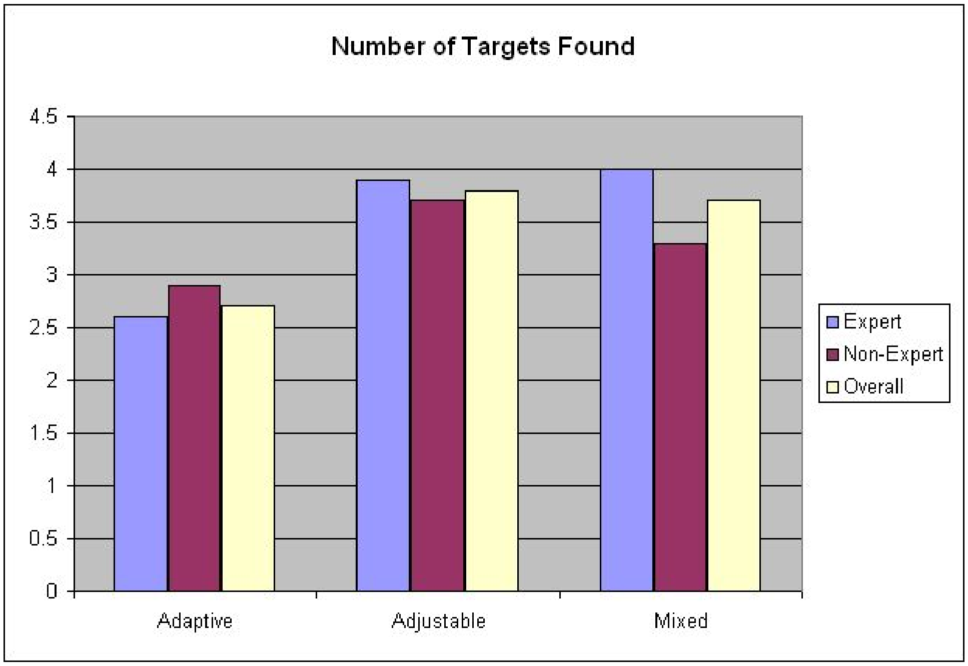


Figure 4.5: Total number of missing people found

The most ecologically valid measure is the number of missing people that were found; measured results were found to be statistically significant with an F-value of 7.37 and p-value of 0.0014. As described earlier, each scenario actually involved five missing people, to allow for averaging the time to discovery and discovering exactly how many missing people a person could find in the given time period. Figure 4.5 shows how adjustable autonomy and mixed initiative resulted in the most missing people found on average, while adaptive autonomy did significantly worse. This poor performance by adaptive autonomy can be directly related to the agents' lack of expert knowledge and poor decision making in their choice of which areas to search; both factors which could be eliminated by allowing an expert to input their knowledge. It must be noted that while the difference between adaptive and adjustable autonomy was statistically significant, the difference between adjustable autonomy and mixed initiative was not statistically significant (with expert supervisor, t-value was 0.17 and p-value was 0.05; with non-expert supervisor, t-value was 0.4 and p-value was 0.05), and the difference between mixed initiative and adaptive autonomy was only significant with an expert supervisor (t-value of 2.97 and p-value of 0.05).

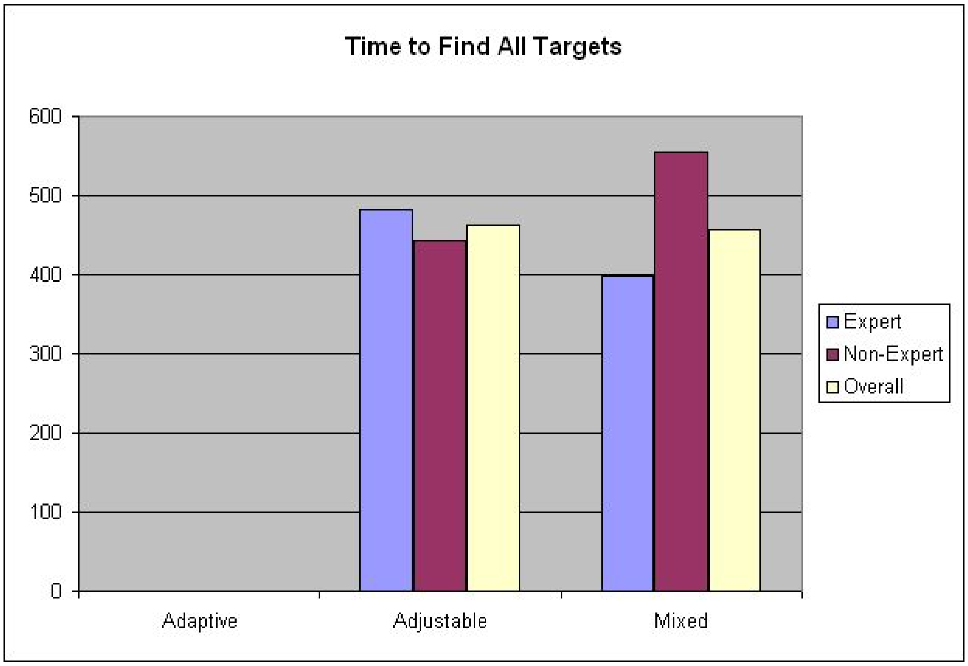


Figure 4.6: The time to find all five missing people, when applicable

Looking at just the scenarios where the supervisor was able to find all five missing people, we see that mixed initiative required less time than adjustable autonomy when the supervisor was an expert, while adjustable autonomy required the least amount of time when the supervisor was a non-expert (Figure 4.6). When agents were operating under mixed initiative, the supervisor would often override the agent's choice of actions with one they deemed more effective. This often resulted in a waste of the agent's time and travel, but if the supervisor possessed expert knowledge, this overriding resulted in a better end-performance. If the supervisor did not possess expert knowledge, the waste was often not worth the change in actions. Using adjustable autonomy, agents never took initiative so there was little difference between expert and non-expert scenarios. Because the feat of finding all five missing people was accomplished so infrequently, these results are not statistically significant (F/p value across expert level of 0.9/2.79, F/p value across autonomy type of 0.01/0.922, F/p value of two-way interaction of 2.79/0.1207), but may serve as an anecdotal indication to guide future studies.

4.2.2 Secondary Task Performance

Secondary tasks involved finding and classifying backpack and distracting items. Metrics which indirectly affected task performance such as the average searcher distance and the percent of terrain coverage also fell into this category.

Backpack and Distracting Items

Overall, agents using adaptive autonomy allowed the supervisor to classify the most items (Figure 4.7). This is statistically significant with an F-value of 28.0 and p-value of 0.0001. However, this metric does not tell the whole story. 88 percent (400 out of 455) of the items were distracting items. Adaptive autonomy's strength was covering a large amount of terrain effectively and quickly, allowing it to find many of the items, with a proportionally high ratio of distracting items to backpack items. With a supervisor participating in mixed initiative and adjustable autonomy, coverage was less inclusive but focused on areas with high probability of containing the missing person and their backpack items. This enabled mixed initiative and adjustable autonomy to out-perform adaptive autonomy when just backpack items were considered (Figure 4.8). This is also statistically significant with an F-value of 4.55 and p-value of 0.0143. The lone exception was the difference between adjustable autonomy and mixed initiative, which was not significant (t-value of 0.62, p-value of 0.05).

This trend holds when we consider the time needed to find the first backpack item on the map, the first backpack item from each of four of the five missing people, and the first backpack item from each of the five missing people (Figure 4.9). This was assuming that one, four, and five missing people were found, respectively. These results were also statistically significant; F- and p-values can be found in Appendix D. Adaptive autonomy was quicker to find the backpack items; the only difference that was not statistically significant was between adjustable autonomy and mixed initiative with a non-expert supervisor (t-value of 0.31, p-value of 0.05).

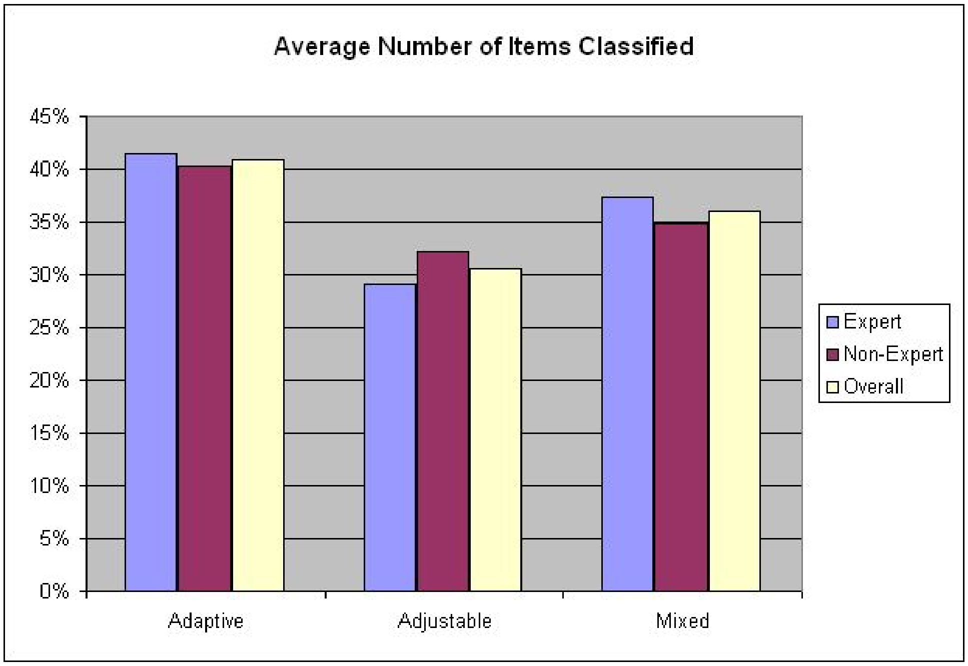


Figure 4.7: Average percent of items classified

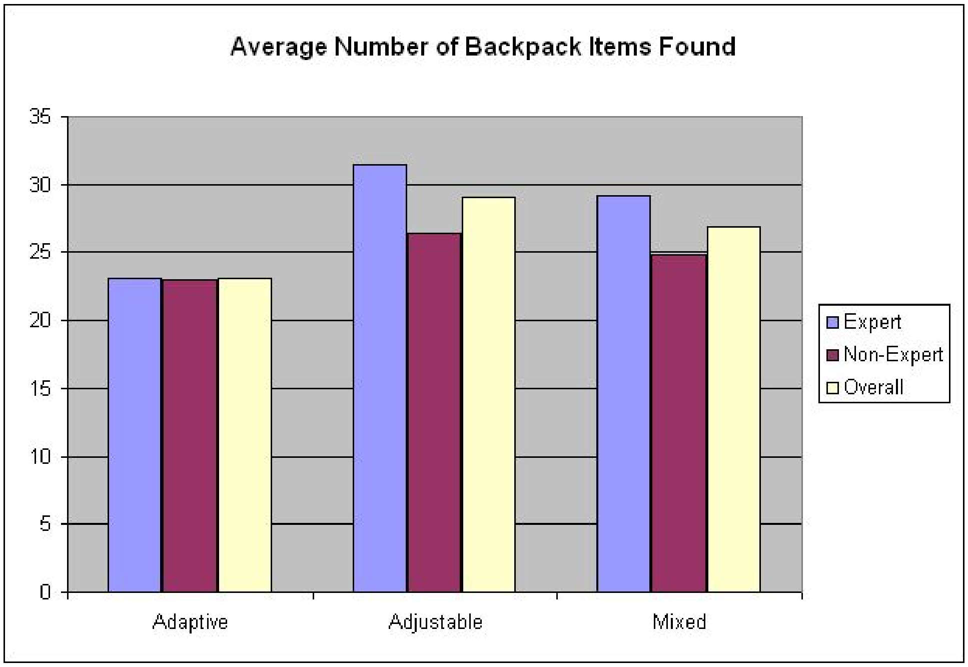


Figure 4.8: Average number of backpack items found

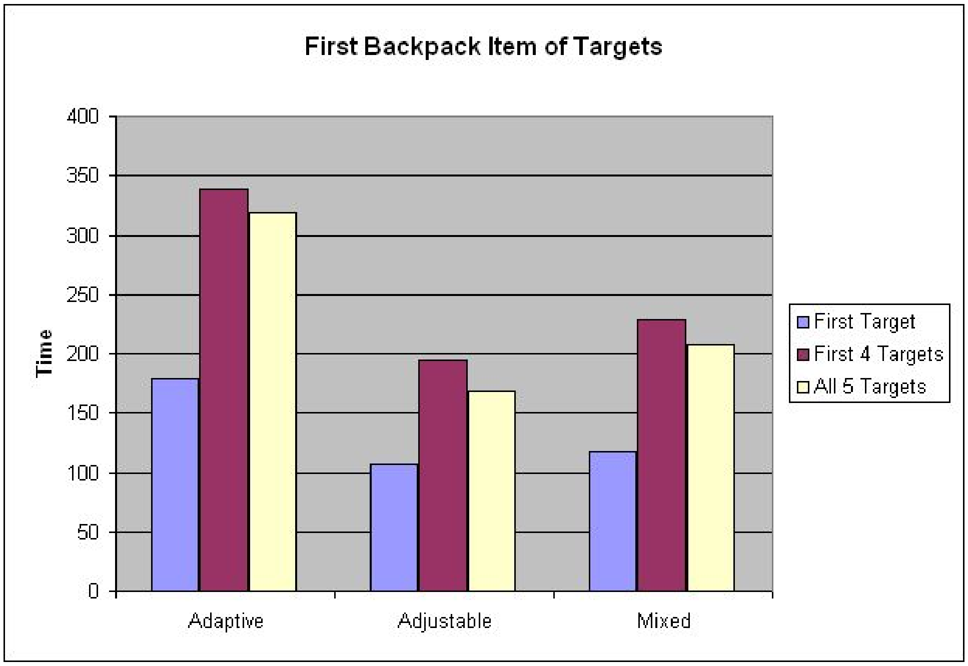


Figure 4.9: Average time to find at least one item from various numbers of missing people

Distance and Coverage

Metrics that fall under distance and area coverage are very interconnected. Although these metrics can be difficult to interpret, they can also provide insight into why the different types of autonomy worked better than others in different scenarios.

Adaptive autonomy and mixed initiative resulted in the most distance covered by the searchers (Figure 4.10). In both cases, the searchers had the autonomy to immediately choose a search area if they would otherwise become idle, so their performance in this metric was almost identical. Adjustable autonomy performed on average 15 percent worse in this metric than adaptive autonomy. When agents at a medium or low level of autonomy completed a task, they would stop moving to await further instructions, reducing their possible searcher distance. These differences were statistically significant (F-value of 54.12, p-value of 0.0001).

It must be noted that average searcher distance can be misleading when applied to live WiSAR scenarios. Different types of searchers will move at different speeds (e.g. ground searchers will move slower than air searchers, and horseback riders may move at different speed than ATV riders), and the type of terrain will further affect travel. These are all

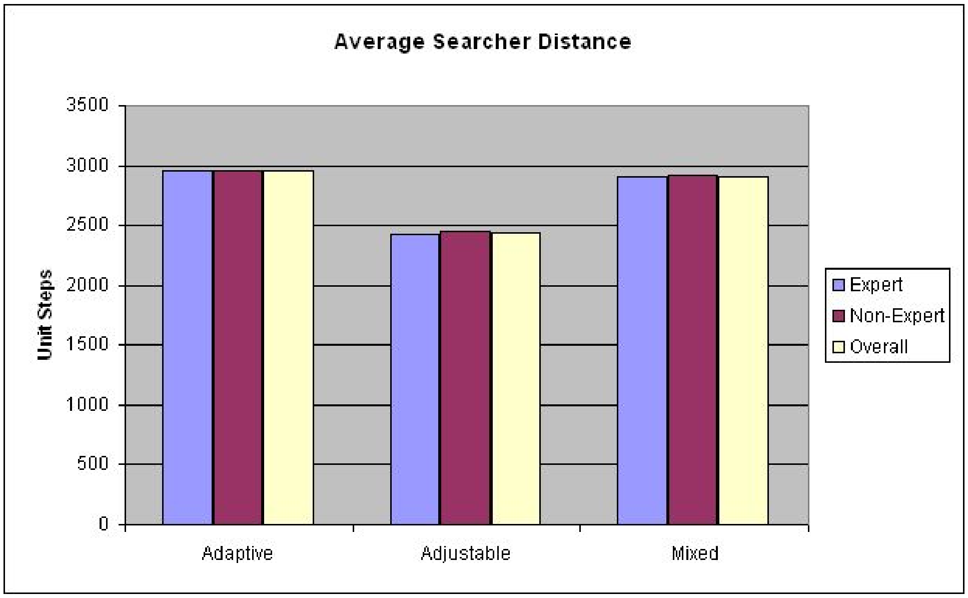


Figure 4.10: Average searcher distance

factors that would affect the average searcher distance, and were not taken into account by our simulator. The metric does show us, however, that adaptive autonomy and mixed initiative can result in more travel, which may loosely be translated as giving a better chance of discovery of the victim over agents employing adjustable autonomy.

Coverage is related to distance traveled, and is expressed as a percentage of the available area since the four different scenarios had different amounts of accessible terrain. Figure 4.11 shows that adaptive autonomy resulted in the best coverage, followed by mixed initiative and adjustable autonomy, in that order. These results were statistically significant with an F-value of 52.61 and p-value of 0.0001.

In addition, the differences between expert and non-expert supervisors were statistically significant (F-value of 7.88 and p-value of 0.0067) for adjustable autonomy (t-value of 12.18 and p-value of 0.05) and mixed initiative (t-value of 3.08 and p-value of 0.05). Under both these types of autonomy, expert supervisors would focus their search on areas their "expert knowledge" indicated had the highest probability of containing the missing people, significantly reducing the amount of area they covered (while keeping their travel distance high).

In other words, coverage is not necessarily proportional to searcher distance. Since

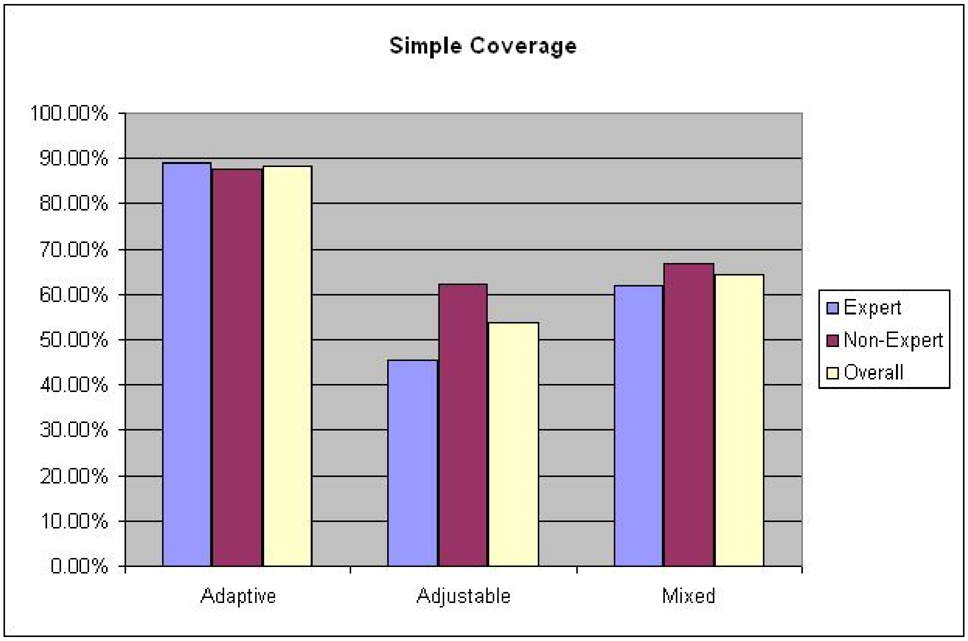


Figure 4.11: Average coverage

the searchers operating under mixed initiative traveled as much distance as the searchers using adaptive autonomy, the difference in the amount of coverage is due to the searchers operating under mixed initiative repeatedly covering the same ground. This is not necessarily bad, since a single pass over a given search area does not result in a 100 percent probability of detection, and some areas have a higher probability of containing the missing person than others. In other words, redundancy can increase our probability of detection. Redundancy is also reflected in the probability of success, where repeatedly searching in areas of high probability of containing the missing person can lead to a high probability of success while still result in a low coverage value.

When broken down by terrain type, adaptive autonomy did far better on simple terrains than on complex ones. The agents operating under adaptive autonomy did not have the same level of intelligence as most human supervisors when it came to path planning or choosing their next search area, resulting in more time spent traveling on complex terrains and less time searching. Agents operating under adjustable autonomy or mixed initiative saw a far smaller difference in performance between complex and simple terrains (Figure 4.12). In future work that attempts to combine the different types of autonomy to take

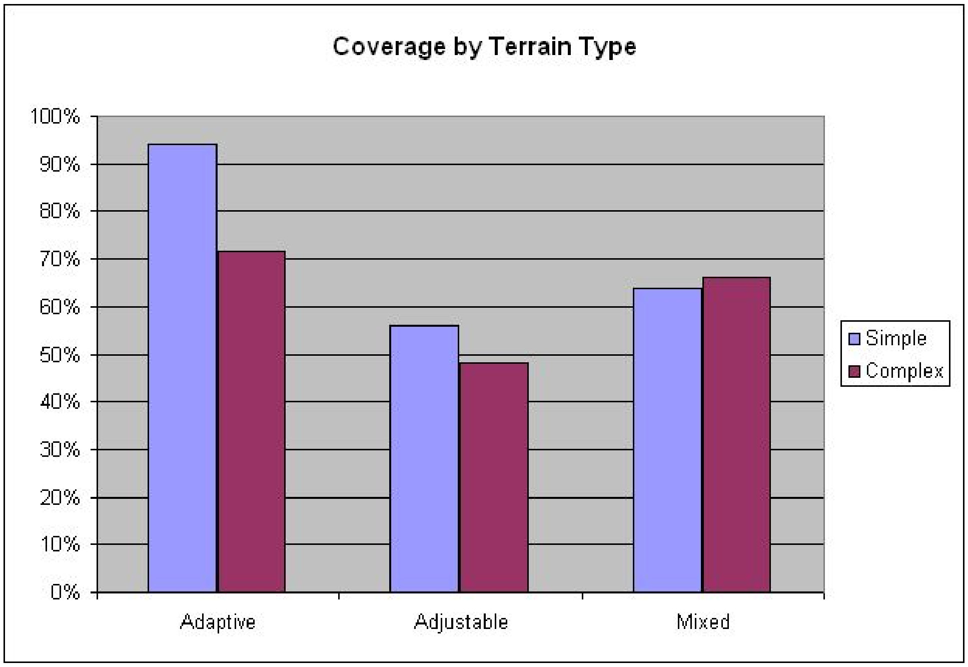


Figure 4.12: Average coverage by terrain type

advantage of each of their strengths, specifically adaptive autonomy and mixed initiative, this might be a key area to consider.

4.2.3 Management Overhead/Workload

Management overhead and workload metrics look at the performance of the supervisor. Although they do not directly measure task performance, they are connected and as a result, cannot be ignored. Decreasing management overhead or workload can improve task performance, and increasing overhead or workload can negatively affect task performance.

Management Overhead

To gives us an initial insight into the amount of management overhead, we can look at the amount of time that agents sit idle. Although searchers operating under adaptive autonomy and mixed initiative are never inactive by design, adjustable autonomy results in the searchers sitting idle 18 percent of the time. This was due to high operator workload (the supervisor unable to respond), a lack of supervisor awareness when agents finished a task (the supervisor unaware of a need to respond), and difficulty in finding the inactive

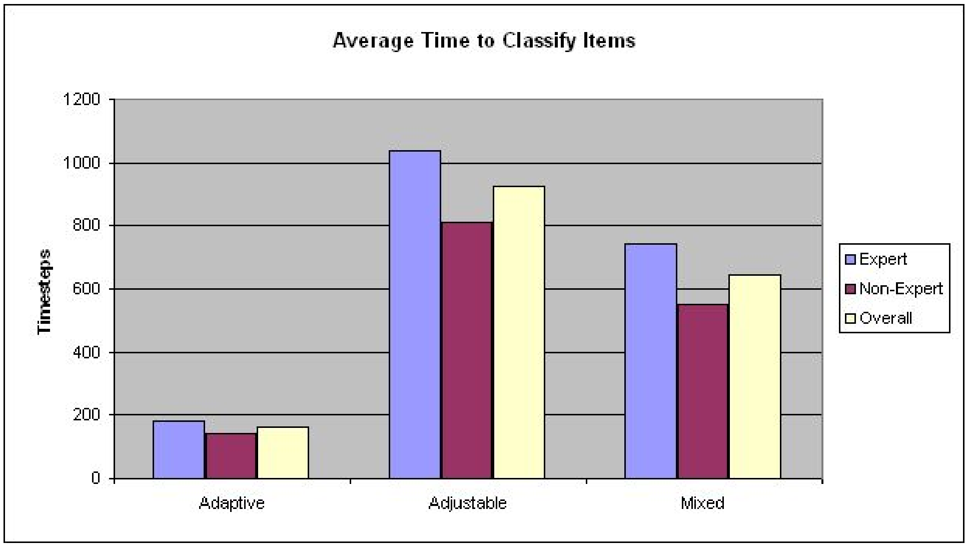


Figure 4.13: Average time to classify items

workers because of interface design or other reasons.

We can also look at the length of time it took supervisors to classify items that were found. As expected, speed of classifying was roughly inversely proportional to the amount of workload. Adaptive autonomy served as a baseline measurement, since the supervisors had no other workload than classifying items. Speed was correspondingly high, on average 163 timesteps. When searchers were using adjustable autonomy, supervisors had to give them a lot of attention and their performance suffered (923 timesteps per classification). Using mixed initiative, supervisors were able to take advantage of the searchers' increased independence and focus more time on classifying items, resulting in 550 timesteps per classification on average (Figure 4.13). These results were statistically significant (F-value of 5.97, p-value of 0.0043).

One more interesting metric can be considered, the average length of time between finding the first item belonging to a missing person, and finding the person. This takes into account several variables such as response speed and effectiveness of search. Interestingly, adaptive autonomy resulted in the fastest time, with mixed initiative second and adjustable autonomy last (Figure 4.14). These results were statistically significant (F-value of 5.39, p-value of 0.007).

This difference could be explained by three factors. First and possibly most important,

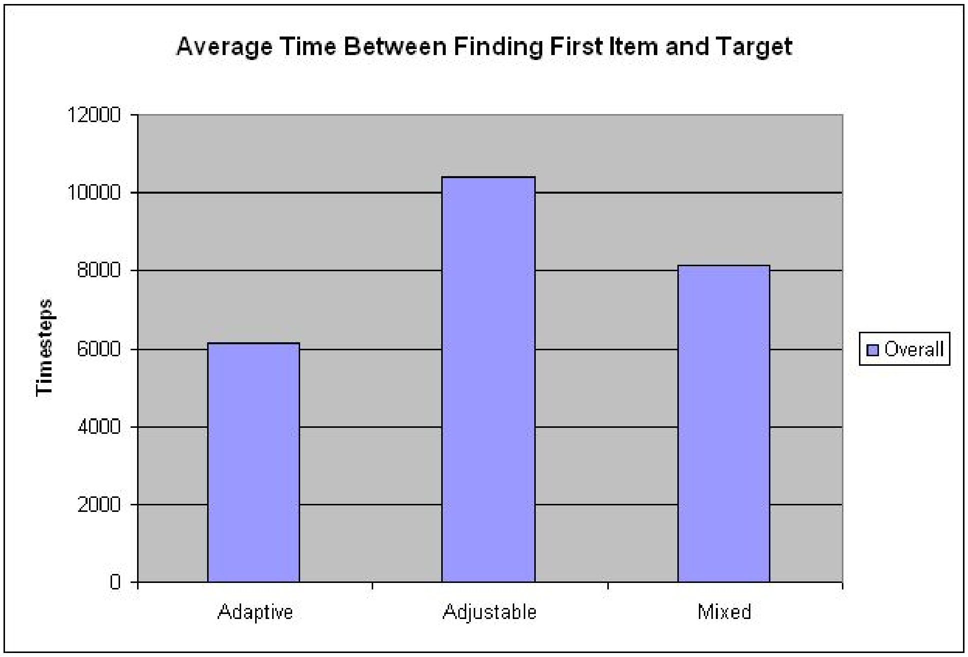


Figure 4.14: Average time between finding first item and missing person

supervisors managing searchers using adjustable autonomy and mixed initiative had the workload of managing the searchers and search areas in addition to classifying items, a task that adaptive autonomy did not require the supervisor to do. As a result, a missing person could languish in the "found items" list far longer than it would under adaptive autonomy before being classified. Two, searchers using adaptive autonomy and mixed initiative reacted instantly to the classification of a backpack item by creating a more specific search area and beginning a detailed search, resulting in a faster discovery than searchers using adjustable autonomy. Three, qualitative observations indicated that the human supervisor frequently became impatient with searchers. If the searchers employed adjustable autonomy or mixed initiative, the supervisors could micromanage them. Each time the supervisor reassigned them, however, the searchers were forced to restart their search, delaying the discovery of the missing person.

Workload

Figure 4.15 shows the supervisor's workload in scenarios using adjustable autonomy compared to scenarios using mixed initiative. Workload in these measurements included se-

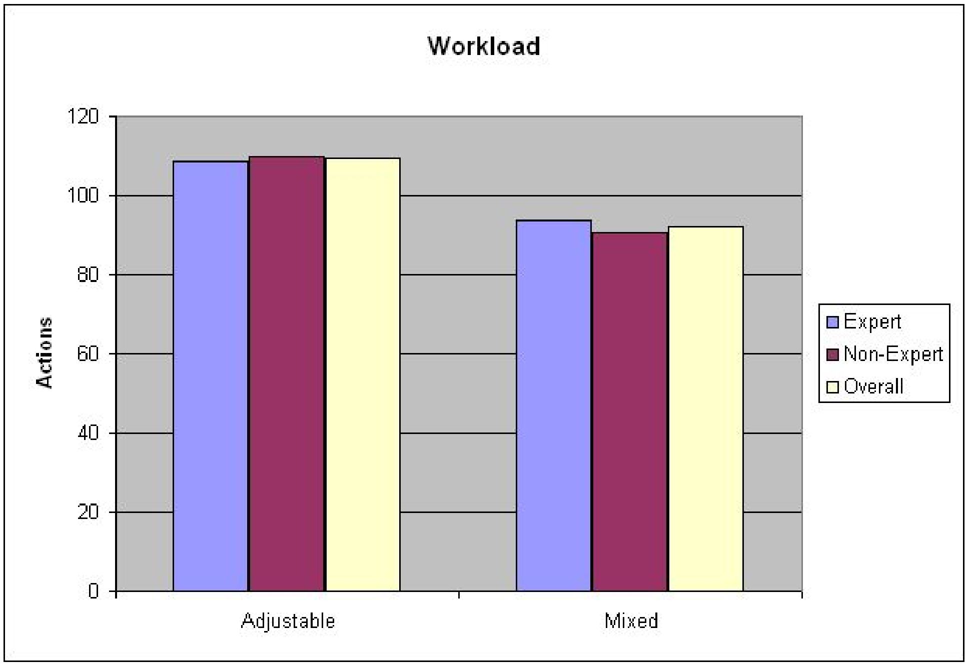


Figure 4.15: Workload

lecting searchers, creating search areas, assigning searchers to a search area, assigning searchers to travel to a point on the map, deleting search areas, and changing searcher autonomy levels. Workload in scenarios using adaptive autonomy is not applicable since supervisors were not given the ability to perform any of the afore-mentioned actions.

We can see that due to the agents' increase in initiative, scenarios where agents employed mixed initiative resulted in less workload for the supervisor. This result is statistically significant (F-value of 54.75, p-value of 0.0001). In addition, this benefit is reflected in such metrics as the time it took supervisors to classify items. With less workload, supervisors were able to devote more time to classifying items, resulting in a shorter average time to classify each item. In a real wilderness search and rescue scenario, increasing searcher autonomy might allow the supervisor more time to manage logistics, direct agents in trouble areas, and talk to media, family, or witnesses.

4.2.4 Discussion

As expected, adaptive autonomy, adjustable autonomy, and mixed initiative each performed well in different metrics. Of note, adjustable autonomy and mixed initiative were strongest in metrics falling under the primary and secondary task performance categories, while adaptive autonomy fared especially well in management overhead/workload metrics. The supervisor's level of expertise made little difference in the results; expert and non-expert supervisors were helped or hindered by the type of autonomy equally in most metrics.

A qualitative observation was made that when mixed initiative or adjustable autonomy was used, most human controllers abandoned an area after finding the target in it. When adaptive autonomy was being used, the agents would complete searching an area regardless of whether or not a missing person had been found in it. This tended to weight the probability of success in favor of adaptive autonomy, even if the overall success metric of number of missing people found showed that mixed initiative or adjustable autonomy performed better. As a result, we might expect that in an experiment designed to measure only probability of success, without searchers actually finding the missing people and therefore short-circuiting their search, mixed initiative and adjustable autonomy would perform even better when compared to adaptive autonomy.

In addition, during the course of the experiments, it was observed that in scenarios using adjustable autonomy or mixed initiative, supervisors frequently and unnecessarily decreased performance of the agents by micromanagement. Often, the supervisors would grow impatient as the searchers searched an area. This would cause the supervisor to delete the search area, forcing the searchers to stop work. The supervisor would then create a new search area that almost exactly matched the original search area, perhaps with one border extended or moved, then reassign the searchers to the new area. This resulted in the searchers not only restarting their search, but largely covering the same ground they had previously covered. Although some redundancy in coverage is desired in areas that have a high probability of containing the missing person, the redundancy described above is a result of impatience and not expert knowledge.

Also, supervisors would often not take travel time into account, assigning agents to search areas a long distance away rather than assigning agents after a spatial assessment of free agents and unassigned search areas. As the agents operating under adaptive autonomy did not take travel time into account either, this resulted in a somewhat uniform degradation of performance, and supervisors in live scenarios are presumably experienced enough to avoid this mistake.

Chapter 5

Conclusions

5.1 Conclusions

Table 5.1 reviews the results from the individual metrics. Note that we have subjectively divided the metrics by relative importance, as outlined in section 3.3.

Scenarios employing mixed initiative resulted in better performance than scenarios where adaptive or adjustable autonomy was used. Although mixed initiative did not rank first in several metrics, it ranked first in more of the metrics than either adaptive or adjustable autonomy, including both of the "High Importance" metrics and many of the "Medium Importance" metrics. In metrics where it did not rank first, it came in second place in many–often a very close second. Of note, adjustable autonomy and mixed initiative were strongest in metrics falling under the primary and secondary task performance categories, while adaptive autonomy fared especially well in management overhead/workload metrics.

This variance of performance suggests that mixed initiative is the best primary autonomy type to use. Searchers operating under mixed initiative have all the initiative of adaptive autonomy, while still giving the supervisor the control of adjustable autonomy. This suggests that it is possible to emulate any of the studied autonomy types (and their performance) using just mixed initiative.

Whether in wilderness search and rescue or another domain, the supervisor can deter-

Table 5.1: Overall Autonomy Rankings

High Importance Metrics	Adaptive	Adjustable	Mixed Initiative
Probability of success		2nd	1st
Number of targets found		1st	1st
Medium Importance Metrics	Adaptive	Adjustable	Mixed Initiative
First backpack item		1st	2nd
Greatest improvement from non-expert to expert		2nd	1st
Time to find all five targets	N/A	1st	1st
Time from first item to target	1st		2nd
Time to classify items	1st		2nd
Number of backpack items classified		1st	2nd
Variance between terrain types		2nd	1st
Workload	N/A	2nd	1st
Low Importance Metrics	Adaptive	Adjustable	Mixed Initiative
Searcher distance	1st		1st
Simple coverage	1st		2nd
Number of items classified	1st		2nd

mine the most important metrics by careful analysis. Following that, the type of autonomy which would result in the best performance in those metrics can be determined. As a result of the agents operating under mixed initiative, if the type of autonomy that results in the best performance is adaptive autonomy, the supervisor can allow the robotic agents to take initiative. If adjustable autonomy, the supervisor can take a more hands-on approach to control. In addition, because of the mixed initiative approach, the supervisor can experiment with the correct level of attention to give. In other words, rather than modifying the type of autonomy on the searchers, a supervisor should modify his/her behavior, inspired by adaptive or adjustable autonomy.

For example, adaptive autonomy was able to locate backpack items faster than adjustable autonomy or mixed initiative (Figure 4.9), but searchers operating under adjustable autonomy or mixed initiative were able to locate more (Figure 4.8). This indicates that supervisors should initially allow agents using mixed initiative to operate at their highest autonomy level, taking advantage of the searchers' speed in creating search areas and assigning themselves to the areas, as well as their minimal idleness time. Once the backpack items are located, the supervisors can use their superior knowledge of the overall situation to assess the best local areas to search and the length of time or depth of search.

5.2 Future Work

Several areas bear more consideration. First, we found that with the relative smallness of the overall search area, most experimenters were able to locate backpack items from most, if not all of the missing people before time ran out. This held true whether adaptive autonomy, adjustable autonomy, or mixed initiative was true. It also had the effect of minimizing the importance of the expert knowledge. Using expert knowledge, the supervisors were able to locate the missing people quickly, but even without the expert knowledge, there was enough time and the search area was small enough that most of the supervisors eventually located most of the missing people anyway. If agents were only capable of searching a much smaller fraction of the total search area due to increased search area size or less time (situations more true to live wilderness search and rescue scenarios), it is our prediction that the expert knowledge would increase in value dramatically.

In addition, there are several areas where variations could be explored. The number and capabilities of the searchers, the various types or levels of autonomy, different terrains, the type and number of secondary tasks, the type and amount of information given the supervisor, etc.

There are several larger areas that bear more discussion, however. In our experiments, we limited our attention to agents employing a supervisor-supervised relationship. Under this circumstance, we were able to show that that mixed initiative can provide better performance over several key metrics. Future experiments may look at situations involving peer-to-peer relationships to see if the results hold. Is one peer possessing expert knowledge able to influence the overall performance to any major degree? How many peers possessing expert knowledge unknown by the remaining peers would it take to affect performance to a given degree?

Although we focused on a homogeneous set of agents (generic “searchers” with a standard set of capabilities), extending to heterogeneous agents would improve ecological validity. Different classes of agents could be given different speeds, varying abilities to communicate, and different task capabilities. For example, unmanned aerial vehicles (UAVs) have the ability to fly, allowing them to not only cover ground faster (albeit with a possibly lower probability of detection), but allowing them to search terrain inaccessible to ground searchers. Even among ground searchers, abilities could differ. Dog teams can follow chemical trails, teams with SCUBA gear could search lakes or waterways, and teams with specialized climbing gear could search cliffs or other mountainous regions. With these differences in capabilities, each class of agents would have different levels of autonomy and events to trigger changes in that level.

Blurring the line between several of the ideas above, introducing imperfect information and task execution could lead to even greater ecological validity. In real life, especially a domain as inexact and error-prone as the “wilderness,” situations are rarely as precise as a simulation. Noise could be introduced into location reporting or other communication between searchers and supervisors, searchers could perform tasks imperfectly or outright incorrectly, the ownership of discovered items could be ambiguous, and the missing person usually has the ability to change their location throughout the duration of the search. Each of these issues could be explored, along with their effect on overall performance.

Appendix A

Workload Data T-Values

This appendix contains the raw data from experiment one, where performance was compared over adaptive autonomy, adjustable autonomy, and mixed initiative; and over high and low workload.

For each metric shown, the first row contains the raw value calculated over all scenarios. The data is displayed in the following format: high workload/low workload/overall or average.

The second row contains the t-values and degrees of freedom for nine selected interactions. The degrees of freedom for each interaction are enclosed in parentheses.

The first three t-values (in the adaptive autonomy column) are adaptive autonomy/high workload compared to adaptive autonomy/low workload; adaptive autonomy/high workload compared to adjustable autonomy/high workload; adaptive autonomy/low workload compared to adjustable autonomy/low workload.

The next three t-values (in the adjustable autonomy column) are adjustable autonomy/high workload compared to adjustable autonomy/low workload; adjustable autonomy/high workload compared to mixed initiative/high workload; adjustable autonomy/low workload compared to mixed initiative/low workload.

The final three t-values (in the adjustable autonomy column) are mixed initiative/high workload compared to mixed initiative/low workload; mixed initiative/high workload compared to adaptive autonomy/high workload; mixed initiative/low workload compared to adaptive autonomy/low workload.

Table A.1: Workload Data and T-Values

Metric Name		
Adaptive Values	Adjustable Values	Mixed Initiative Values
Average time between finding first item and finding target of the item		
6603.3/6061.7/6398.7	7791.8/8591.3/8145.1	7631.8/9374.2/8737.6
16.22(10)/76.91(11)/84.3(9)	1.92(10)/20.54(9)/42.65(12)	59.12(11)/38.44(10)/132.56(11)
Average searcher distance		
2862.1/2889.2/2873.4	2371.8/2412.5/2392.2	2874/2855/2862.3
3.48(10)/51.49(11)/46.49(9)	3.5(10)/59.01(9)/46.61(12)	3.63(11)/1.88(10)/4.93(11)
Average timesteps inactive		
0%/0%/0%	14.7%/18%/16.3%	0%/0%/0%
0(N/A)/10.15(11)/11.2(9)	1.5(10)/10.15(9)/11.2(12)	0(N/A)/0(N/A)/0(N/A)
Average length of time to classify items		
519.5/256/445.2	1726.4/926.3/1432.7	2059/511.3/1275.7
18.68(10)/62.6(11)/66.53(9)	42.89(10)/25.44(9)/31.64(12)	79.43(11)/78.94(10)/30.96(11)
Simple coverage		
98.5%/99.8%/99.1%	80.5%/83.3%/81.9%	89.6%/92.1%/91.1%
1.25(10)/9.01(11)/9.87(9)	1.16(10)/4.09(9)/4.54(12)	1.5(11)/6.06(10)/6.08(11)
Average number of items classified		
76.1%/76.1%/76.1%	59.4%/65.6%/62.5%	62.8%/71.8%/68.4%
0.03(10)/9.2(11)/6.13(9)	3.02(10)/1.67(9)/3.44(12)	5.09(11)/7.66(10)/2.98(11)
Number of targets found		
4/3.4/3.8	4/3.2/3.6	3.8/4.1/4
1.02(10)/0(11)/0.4(9)	1.36(10)/0.29(9)/1.74(12)	0.51(11)/0.31(10)/1.28(11)
Number of times changed autonomy		
306/302.4/304.5	441/639.5/540.3	943.4/735.5/815.5

0.46(10)/17.71(11)/35.74(9)	21.37(10)/46.6(9)/10.78(12)	19.89(11)/62.72(10)/53.03(11)
Number of backpack items found		
37.9/35.2/36.8	34.7/34.7/34.7	34.2/40.1/37.8
2.21(10)/2.19(11)/0.39(9)	0(10)/0.23(9)/3.61(12)	3.02(11)/2(10)/3.52(11)
Number of times found all 5 targets		
0.4/0/0.3	0.5/0/0.3	0.4/0.5/0.5
1.55(10)/0.17(11)/0(N/A)	1.65(10)/0.22(9)/1.93(12)	0.24(11)/0.07(10)/1.93(11)
Workload		
1.2/1/1.1	72.7/90.8/81.8	45.6/54.1/50.8
0.64(8)/30.45(10)/38.75(8)	5.52(10)/10.19(9)/13.47(12)	4.45(11)/34.24(9)/37.08(10)

Appendix B

Expert Data T-Values

This appendix contains the raw data from experiment two, where performance was compared over adaptive autonomy, adjustable autonomy, and mixed initiative; and over expert and non-expert supervisors.

For each metric shown, the first row contains the raw value calculated over all scenarios. The data is displayed in the following format: expert supervisor/non-expert supervisor/overall or average.

The second row contains the t-values and degrees of freedom for nine selected interactions. The degrees of freedom for each interaction are enclosed in parentheses.

The first three t-values (in the adaptive autonomy column) are adaptive autonomy/expert supervisor compared to adaptive autonomy/non-expert supervisor; adaptive autonomy/expert supervisor compared to adjustable autonomy/expert supervisor; adaptive autonomy/non-expert supervisor compared to adjustable autonomy/non-expert supervisor.

The next three t-values (in the adjustable autonomy column) are adjustable autonomy/expert supervisor compared to adjustable autonomy/non-expert supervisor; adjustable autonomy/expert supervisor compared to mixed initiative/expert supervisor; adjustable autonomy/non-expert supervisor compared to mixed initiative/non-expert supervisor.

The final three t-values (in the adjustable autonomy column) are mixed initiative/expert supervisor compared to mixed initiative/non-expert supervisor; mixed initiative/expert supervisor compared to adaptive autonomy/expert supervisor; mixed initiative/non-expert supervisor compared to adaptive autonomy/non-expert supervisor.

Table B.1: Expert Data and T-Values

Metric Name		
Adaptive Values	Adjustable Values	Mixed Initiative Values
Average time between finding first item and finding target of the item		
6125.3/6179.3/6150.7	10347/10487.3/10410.4	8033.3/8211.4/8118.1
30.05(17)/187.74(21)/169.96(19)	36.47(23)/65.96(22)/66.65(21)	24.97(20)/92.06(19)/105.16(18)
Average searcher distance		
2961.7/2954.4/2958.1	2425.8/2445.6/2435.3	2907.4/2910.9/2909.2
4.42(19)/106.3(21)/89.31(21)	2.9(23)/90.34(22)/83.21(21)	1.77(20)/15.92(19)/9.74(20)
Average length of time to classify items		
183.3/142.9/163.4	1038.6/809.2/922.7	744.6/549.9/646.4
12.43(19)/91.99(21)/79.05(21)	19.4(23)/23.44(22)/22.84(21)	14.22(20)/56.94(19)/73.7(20)
Uniform weighted probability of success		
36.5%/35.5%/36%	29.9%/29.6%/29.7%	28%/28.5%/28.2%
1.87(19)/11.32(21)/9.21(21)	0.53(23)/5.33(22)/3.88(21)	0.51(20)/17.14(19)/14.63(20)
Linear weighted probability of success		
47.1%/49.1%/48.1%	53.2%/53.6%/53.4%	57.5%/56.9%/57.2%
2.16(19)/8.18(21)/7.16(21)	0.56(23)/8.41(22)/8.24(21)	1.29(20)/16.15(19)/17.67(20)
Limit weighted probability of success		
49%/50.1%/49.6%	53.4%/53.8%/53.6%	57.7%/57%/57.3%
1.08(19)/6.45(21)/5.81(21)	0.56(23)/8.44(22)/8.26(21)	1.24(20)/14.88(19)/15.18(20)
Exponential weighted probability of success		
47%/48.7%/47.9%	52.7%/53%/52.8%	56.9%/56.4%/56.6%
1.81(19)/7.87(21)/6.94(21)	0.59(23)/8.45(22)/8.25(21)	1.27(20)/16.21(19)/17.2(20)
Sigmoidal weighted probability of success		
47.1%/49.1%/48.1%	53.2%/53.6%/53.4%	57.5%/56.9%/57.2%

2.19(19)/8.15(21)/7.1(21)	0.56(23)/8.39(22)/8.22(21)	1.3(20)/16.03(19)/17.51(20)
Simple coverage		
89.1%/87.6%/88.4%	45.7%/62.5%/53.8%	61.9%/66.8%/64.4%
1.97(19)/33.49(21)/15.87(21)	12.18(23)/12.31(22)/2.54(21)	3.08(20)/19.06(19)/14.14(20)
Simple coverage broken down by terrain		
93.7%/94.3%/94%	48.1%/62.2%/55.9%	62.4%/65.2%/63.9%
0.74(13)/38.77(14)/22.66(15)	8.68(16)/8.95(14)/1.73(17)	1.67(15)/23.07(14)/22.39(14)
Simple coverage broken down by terrain		
70.7%/72%/71.5%	41.9%/63.8%/48.2%	60.5%/71.6%/66.1%
2.04(4)/19.2(5)/6.25(4)	11.7(5)/7.69(6)/4.15(2)	4.55(3)/4.84(3)/0.97(4)
Average number of items classified		
41.4%/40.3%/40.9%	29.1%/32.2%/30.6%	37.4%/34.9%/36.1%
1.89(19)/14.06(21)/7.93(21)	3.8(23)/10.03(22)/3.29(21)	2.63(20)/4.46(19)/4.75(20)
Number of targets found		
2.6/2.9/2.7	3.9/3.7/3.8	4/3.3/3.7
0.14(17)/3.06(21)/2.31(18)	0.46(22)/0.17(22)/0.4(20)	0.96(20)/2.97(19)/1.87(18)
Number of backpack items found		
23.1/23/23.1	31.5/26.4/29	29.2/24.8/26.9
0.79(19)/8.78(21)/3.56(21)	4.68(23)/2.18(22)/0.62(21)	2.87(20)/5.4(19)/2.89(20)
Time to find all 5 targets		
(N/A)/(N/A)/(N/A)	483.3/443.3/463.3	397.4/554.7/456.4
149.63(2)/72.57(2)	5.79(6)/13.55(7)/16.59(5)	25.68(6)/72.94(3)/198.35(1)
Number of times found all 5 targets		
0/0/0	0.3/0.3/0.3	0.5/0.3/0.3
0(N/A)/1.6(21)/1.65(19)	0.09(23)/0.51(22)/0.21(21)	0.61(20)/2.09(19)/1.32(18)
Workload		
1.2/2.4/1.8	108.8/110/109.4	93.8/90.7/92.2
1.64(19)/68.81(21)/47.43(21)	0.43(23)/5.68(22)/3.97(21)	1.79(20)/42.98(19)/51.19(20)

Average time to find first backpack item		
173.2/185.6/179.4 6.37(19)/30.51(21)/26.15(21)	98.5/115.4/106.6 7.9(23)/2.49(22)/0.31(21)	105.4/129.5/118 3.37(20)/21.21(19)/23.76(20)
Average time to find all 5 first backpack items		
336/341.3/338.6 5.65(19)/57.26(21)/34.98(21)	171.5/222.1/194.7 16.45(23)/7.15(22)/6.35(21)	193/264.2/228.9 12.9(20)/36.63(19)/28.94(20)
Average time to find first 4 backpack items		
311.4/325.4/318.4 7.6(19)/59.38(21)/36.32(21)	142.7/197.6/168.5 19.3(23)/7.93(22)/5.08(21)	170.6/242.2/208 12.41(20)/36.51(19)/30.59(20)

Appendix C

Workload Data F-Values

This appendix contains the F- and p-values resulting from an analysis of variance (ANOVA) on data from experiment one, where performance was compared over adaptive autonomy, adjustable autonomy, and mixed initiative; and over high and low workload.

For each metric shown, the first row contains the F-value and the second row contains the p-value. The degrees of freedom in the numerator is two, and the degrees of freedom in the denominator is 31.

Table C.1: Workload Data and F/p-Values

Difference Across Autonomy	Difference Across Workload	Two-Way Interaction
Average time between finding first item and finding target of the item		
0.77	2.22	0.26
0.387	0.1256	0.7727
Average searcher distance		
0.07	13.29	0.03
0.7931	0.0001	0.9705
Average timesteps inactive		
0.08	16.87	0.2
0.7792	.0001	0.8198

Average length of time to classify items		
5.04	2.68 2.38	
0.032	0.0844	0.1093
Simple coverage		
0.11	6.5	0.12
0.7424	0.0044	0.8873
Average number of items classified		
1.62	5.6	1.04
0.2126	0.0084	0.3655
Number of targets found		
0.7	0.43	1
0.4092	0.6543	0.3794
Number of times changed autonomy		
0.57	12.66	1.66
0.456	0.0001	0.2066
Number of backpack items found		
0.24	0.47	0.82
0.6277	0.6294	0.4498
Number of times found all 5 targets		
2.47	0.92	1.96
0.1262	0.4091	0.1579

Appendix D

Expert Data F-Values

This appendix contains the F- and p-values resulting from an analysis of variance (ANOVA) on data from experiment two, where performance was compared over adaptive autonomy, adjustable autonomy, and mixed initiative; and over expert and inexpert supervisors.

For each metric shown, the first row contains the F-value and the second row contains the p-value. The degrees of freedom in the numerator is two, and the degrees of freedom in the denominator is 62.

Table D.1: Expert Data and F/p-Values

Difference Across Autonomy	Difference Across Workload	Two-Way Interaction
Average time between finding first item and finding target of the item		
0.5	5.39	0.02
0.4822	0.007	0.9802
Average searcher distance		
0.02	54.12	0.07
0.888	.0001	0.9325
Average length of time to classify items		
0.67	5.97	0.12
0.4163	0.0043	0.8871

Simple coverage		
7.88	52.61	3.48
0.0067	.0001	0.037
Average number of items classified		
0.01	28	2.67
0.9207	.0001	0.0772
Number of targets found		
0.53	7.37	0.28
0.4695	0.0014	0.7568
Number of backpack items found		
3.76	4.55	0.28
0.057	0.0143	0.7567
Number of times found all 5 targets		
0.17	4.61	0.46
0.6816	0.0137	0.6335
Probability of Success		
1.13	89.53	0.84
0.2919	.0001	0.4366
Workload		
0.01	54.75	0.05
0.9207	.0001	0.9513
Average time to find first backpack item		
2.11	17.47	0.3
0.1515	.0001	0.7419
Average time to find all 5 first backpack items		
6.57	25.87	0.29
0.0129	.0001	0.7493
Average time to find all 5 targets		

0.9	0.01	2.79
2.79	0.922	0.1207
Average time to find first 4 backpack items		
7.11	26.21	0.3
0.0098	.0001	0.7419

Bibliography

[1] R. C. Arkin and T. Balch. *Cooperative multiagent robotic systems.* MIT Press, 1998. Provides a brief overview of the purpose for multiagent robotic systems, and describes schema-based reactive control.

[2] M. Asada and H. Kitano. Springer, 1999. Gives an overview of the second Robot Soccer World Cup, describes the teams, and most applicable: presents a set of related technical papers.

[3] L. Bainbridge. Ironies of automation. In *New Technology and Human Error*, page 271283, 1987. Shows how automation may in many cases actually make an operator's job more difficult or cause other problems.

[4] T. Balch and M. Hybinette. Social potentials for scalable multi-robot formations. In *Robotics and Automation*, volume 1, pages 73–80, 2000. Describes a potential-field method of organizing multiagent robotic teams into various formations, coins the phrase "social potentials" to describe these potential functions, and gives examples of several functions.

[5] K. Barber, A. Goel, and C. Martin. The motivation for dynamic adaptive autonomy in agent-based systems. In *Proceedings of the 1st Asia-Pacific Conference on IAT*, pages 131–140, 1999. Describes dynamic adaptive autonomy and introduces a framework for problem-solving.

[6] K. Barber, A. Goel, and C. Martin. Dynamic adaptive autonomy in multi-agent systems. In *The Journal of Experimental and Theoretical Artificial Intelligence, Special*

Issue on Autonomy Control Software, volume 12, pages 129–147, 2000. Claims agent adaptability as neccessary for effective performance, and describes a decision-making framework for multiagent teams.

[7] K. Barber and C. Martin. Agent autonomy: Specification, measurement, and dynamic adjustment. In *Autonomy Control Software Workshop at Autonomous Agents*, pages 8–15, 1999. Presents a formal definition of agent autonomy.

[8] E. Bonabeau and C. Meyer. Swarm intelligence. In *Harvard Business Review*, pages 107–114, 2001. Describes how the swarming behavior of social insects such as ants and bees can be applied to solve real-life problems.

[9] F. Bourgault, T. Furukawa, , and H. Durrant-Whyte. Coordinated decentralized search for a lost target in a bayesian world. In *Intelligent Robots and Systems*, volume 1, pages 48– 53, 2003. Suggests a technique for locating non-moving targets which involves each search agent creating a Bayesian representation of the target's state.

[10] J. Brookshire. Enhancing multi-robot coordinated teams with sliding autonomy, May 2004. Describes the benefits and possible applications of sliding autonomy in multiagent robotic teams.

[11] J. Brookshire, S. Singh, and R. Simmons. Preliminary results in sliding autonomy for coordinated teams. In *The 2004 Spring Symposium Series*, 2004. Proposes and evaluates a framework which allows an operator to interact with a small heterogenous team of robotic agents. Applies it to a team of structure-building robots.

[12] A. Collinot, A. Drogoul, and P. Benhamou. Agent oriented design of a soccer robot team. In *Multi-Agent Systems*, pages 41–47, 1996. Gives a method of dissecting a global task specification with the intent of assigning subtasks to individual agents, and applies this to a robotic soccer team.

[13] D. Cooper. University of Edinburgh, 2005. An in-depth review of search and rescue, covering all aspects from prevention to preparation to management.

[14] H. Cottam. University of Edinburgh, 1996. Discusses knowledge aquisition and the creation of knowledge base systems in a search and rescue domain.

[15] J. Crandall, M. Goodrich, D. Olsen, and C. Nielsen. Validating human-robot interaction schemes in multi-tasking environments. In *Systems, Man, and Cybernetics-Part A: Systems and Humans*, volume 33, page 325336, 2003. Presents the idea of neglect tolerance, including how to use it to identify the span of control, posssible robotic teams, and the expected performance of the teams.

[16] R. de Boer and J. R. Kok. The incremental development of a synthetic multi-agent system: The uva trilearn 2001 robotic soccer simulation team. In *Master's thesis*, 2002. Describes the development of a robotic soccer team.

[17] T. Drabek. Managing the emergency response. In *Public Administration Review*, volume 45, pages 85–92, 1985. Discusses the inherent social and response structure underlying any emergency management system, including search and rescue.

[18] J. Draper. Teleoperators for advanced manufacturing: Applications and human factors challenges. In *Human Factors and Manufacturing*, volume 5, pages 53–85, 1999. Presents a formal definition of agent autonomy.

[19] G. Dudek, M. Jenkin, E.E. Milios, and D. Wilkes. A taxonomy for multi-agent robotics. In *Autonomous Robots*, volume 3, pages 375–397, 1996. Presents a taxonomy for classifying multiagent systems to allow for structured design decisions, and gives an overview of how existing multiagent systems fit into the taxonomy.

[20] T. Fong, A. Steindfeld, D. Kaber, M. Lewis, J. Scholtz, A. Schultz, and M. A. Goodrich. Common metrics for human-robot interaction. In *Human Robot Interaction '06*, 2006. This paper outlines possible metrics for use in the field of human-robot interactions, as well as discussing the need for them.

[21] T. Fong, C. Thorpe, and C. Baur. Collaborative control: a robot-centric model for vehicle teleoperation, January 1998. Describes a robot-centered control for teleoperated vehicles, suitable for far-distant agents such as planetary rovers or collaborative control between robotic agents.

[22] M. A. Goodrich, D. Olsen Jr., J. Crandall, and T. Palmer. Experiments in adjustable autonomy. In *Proceedings of the IJCAI 01 Workshop on Autonomy, Delegation, and Control: Interacting with Autonomous Agents*, 2001. Describes the use of adjustable autonomy on agents performing various tasks, also presents the idea of neglect as a metric.

[23] J. Hackman. Humans, robots, and teams, 2007. Describes the concept of levels of autonomy as applied to teams, rather than individual agents.

[24] S. Hart and L. Staveland. Development of nasa-tlx (task load index): results of empirical and theoretical research. In *Human Mental Workload*, pages 139–183, 1988. Presents a way to subjectively measure the workload on system operators.

[25] F. Heger, L. Hiatt, B. Sellner, R. Simmons, and S. Singh. Results in sliding autonomy for multi-robot spatial assembly. In *8th International Symposium on Artificial Intelligence, Robotics and Automation in Space (iSAIRAS)*, 2005. Describes the use of sliding autonomy in multiagent robotic teams while assembling orbital or planetary structures.

[26] J. Hoc and S. Debernard. Respective demands of task and function allocation on human-machine co-operation design: a psychological approach. In *Connection Science*, volume 14, pages 283–295, 2002. Discusses the dynamic allocation of tasks between a human supervisor and robotic agents.

[27] A. Jacoff, E. Messina, and J. Evans. A standard test course for urban search and rescue robots. In *Performance Metrics for Intelligent Systems Workshop*, 2000. An overview of the development of a standard test arena for urban search and rescue.

[28] A. Jacoff, E. Messina, and J. Evans. Experiences in deploying test arenas for autonomous mobile robots. In *NIST Special Publication*, pages 87–94, 2002. Describes a set of reference areans for use in robotic urban search and rescue test. This will allow for structured comparisions bewteen teams of robotics and between different competitions.

[29] D. Olsen Jr. and S. B. Wood. Fan-out: Measuring human control of multiple robots. In *Human factors in computing systems*, pages 231 – 238, 2004. Describes fan-out as a metric, including how it is affected by neglect tolerance and interaction time. Also describes the idea of interaction effort and how to use it as a measure of interaction design effectiveness.

[30] D. Kaber and M. Endsley. The effects of level of automation and adaptive automation on human performance, situation awareness and workload in a dynamic control task. In *Ergonomics*, volume 42, pages 462–492, 1999. Describes the use of levels of automation to manage operator situation awareness, and describes how the use of adaptive autonomation can reduce an operator's workload.

[31] M. Mataric. Minimizing complexity in controlling a mobile robot population. In *Robotics and Automation*, volume 1, pages 830–835, 1992. Describes the tradeoffs of using a peer-to-peer method of task distribution over a set of mobile robots, compared to a hierarchical distribution.

[32] C. Miller and R. Parasuraman. Beyond levels of automation: An architecture for more flexible human-automation collaboration. In *Human Factors and Ergonomics Society 47th Annual Meeting*, pages 182–186, 2003. Broadens the idea of levels of automation, creating a more detailed and flexible break-down over task models.

[33] F. Mondada, L. M. Gambardella, D. Floreano, S. Nolfi, J.-L. Deneubourg, and M. Dorigo. The cooperation of swarm-bots: physical interactions in collective robotics. In *IEEE Robotics and Automation Magazine*, volume 12, pages 21–28, 2005. Discusses

the design and implementation of swarm robotics, and presents a physics-based simulator.

[34] R. Murphy. Marsupial and shape-shifting robots for urban search and rescue. In *Intelligent Systems*, volume 15, pages 14–19, 2000. Describes the use of marsupial and shape-changin robots to facilitate victim recovery and damage assessment, while reducing risk to human searchers in an urban search and rescue scenario.

[35] F. Di Nocera, B. Lorenz, and R. Parasuraman. Consequences of shifting from one level of automation to another: main effects and their stability. In *Human Factors in Design, Safety, and Management*, pages 363–376, 2005. Experimented with the effect of lag on performance when adjustably changing the level of autonomation of agents in a simulated space mission.

[36] J. Odell, H. Parunak, and B. Bauer. Representing agent interaction protocols in uml. In *Agent-Oriented Software Engineering: First International Workshop*, pages 121–139, 2000. Describes how to model autonomous agent interaction in UML, including risk-reduction strategies.

[37] D. Olsen and M.A. Goodrich. Metrics for evaluating human-robot interactions. In *Proceedings of PERMIS*, 2003. Discusses a variety of metrics for measuring a human-robot interface, including free time and fan-out.

[38] R. Parasuraman, T. Sheridan, and C. Wickens. A model for types and levels of human interaction with automation. In *Systems, Man, and Cybernetics*, volume 30, pages 286–297, 2000. Outlines a level of automation model to facilitate the division of tasks between robotic agents and human operators.

[39] D. Perzanowski, A. Schultz, W. Adams, and E. Marsh. Goaltracking in a natural language interface: Towards achieving adjustable autonomy. In *Computational Intelligence in Robotics and Automation*, page 208213, 1999. Presents the idea of context

predicates to keep track of a conversation between a human and robotic agent, facilitating adjustable autonomy and therfore performance of the agent.

[40] T. Setnicka and K. Andrasko. Appalachian Mountain Club, 1980. An overview of wilderness search and rescue, including many live examples.

[41] T. Sheridan. MIT Press, 1992. An early description of the concept of levels of autonomation, defining them as giving or asking for permission, or informing of the decision made.

[42] T. Sheridan and W. Verplank. Human and computer control of undersea teleoperators. In *MIT Man-Machine Laboratory*, 1978. Describes teleoperation, and one of the earliest introductions to the concept of levels of autonomation.

[43] D. Simons and D. Levin. Change blindness. In *Trends in Cognitive Sciences*, volume 1, page 261267, 1997. Overviews the field of change blindness and makes the claim that little information is carried from one view to the next.

[44] L. Urwick. The manager's span of control. In *Harvard Business Review*, volume 34, pages 39–47, 1956. An early description of span of control and fanout, with examples drawn from World War II military situations.

[45] L. Vig and J. Adams. Multi-robot coalition formation. In *IEEE Transactions on Robotics*, volume 22, pages 637–649, 2006. Presents an overview of robotic coalition formation algorithms, as well as uncovering the difficulties of applying the currently-theoretical algorithms to live robots rather than simulated ones.