On Using Mixed-Initiative Control: A Perspective for Managing Large-Scale Robotic Teams

Benjamin Hardin
Computer Science Department
Brigham Young University
Provo, UT USA
bch36@byu.edu

Michael A. Goodrich
Computer Science Department
Brigham Young University
Provo, UT USA
mike@cs.byu.edu

ABSTRACT

Prior work suggests that the potential benefits of mixed initiative management of multiple robots are mitigated by situational factors, including workload and operator expertise. In this paper, we present an experiment where allowing a supervisor and group of searchers to jointly decide the correct level of autonomy for a given situation ("mixed initiative") results in better overall performance than giving an agent exclusive control over their level of autonomy ("adaptive autonomy") or giving a supervisor exclusive control over the agent's level of autonomy ("adjustable autonomy"), regardless of the supervisor's expertise or workload. In light of prior work, we identify two elements of our experiment that appear to be requirements for effective mixed initiative control of large-scale robotic teams: (a) Agents must be capable of making progress toward a goal without having to wait for human input in most circumstances. (b) The operator control interface must help the human to rapidly understand and modify the progress and intent of several agents.

Categories and Subject Descriptors

I.2.9 [**Artificial Intelligence**]: Robotics—*autonomous vehicles, operator interfaces*

General Terms

Human Factors

Keywords

Adaptive Autonomy, Adjustable Autonomy, Mixed Initiative, Human-Robot Interaction, Unmanned Vehicles, User Study

1. INTRODUCTION

Many tasks benefit from or require large numbers of robotic agents, but as the number of agents or teams increases so does the operator's workload. To allow large robot teams[1] without exceeding workload limits, a common solution is to give the agents autonomy. Although autonomy can be characterized in many different ways [10], we operationally define *autonomy* as the set of *tasks* the agents take or are assigned [15]. A set of tasks defines a *role*, so an agent's *level of autonomy* is the agent's role or set of roles [11, 20].

The level of autonomy has the potential to change over time. Three approaches can be taken in dynamically controlling the level of autonomy: *adaptive autonomy* (DA), giving the agent exclusive control; *adjustable autonomy* (JA), giving the supervisor exclusive control; and *mixed initiative* (MI), where the agent and supervisor collaborate to maintain the best perceived level of autonomy.

While a supervisor may have better task-specific knowledge of the situation, the agents may be able to better perform systematic routines. Thus, MI control should theoretically be superior to both JA and DA since MI control better uses the complementary abilities of humans and agents. Unfortunately, some results from the literature suggest that MI control may fail under some circumstances.

In this paper, we present an experiment that compares MI, DA and JA control in a simulated Wilderness Search and Rescue (WiSAR) domain. As implemented in this paper, the human and agents have complementary abilities: the human manager focuses resources better on likely search areas, but the robotic agents tend to more effectively utilize large numbers of agents to systematically cover ground. Since experimental results indicate that MI outperforms DA and JA in this domain, we will identify elements of the interface and autonomy design that appear to be essential for effective mixed initiative control of large-scale robotic teams. Identifying these essential design elements from the specific experiment results offer some lessons that may generalize to other problem domains, though future work should more carefully explore how general these lessons are.

2. RELATED LITERATURE

Using MI in team settings has been studied previously for a single human and a single robot [13], for two or three robots managed by an equal number of supervisors [17], and for a dozen agents operating under swarm behavior [3]. In addition, the impact of DA and JA control on performance and workload has been studied, both for single agents and for teams of robots. [1, 11, 7, 18]. Results indicate that MI control can sometimes fail to achieve its theoretical benefit,

[1] We restrict attention to teams of semi-independent robots rather than swarms.

HRI'09, March 11–13, 2009, La Jolla, California, USA.
Copyright 2009 ACM 978-1-60558-404-1/09/03 ...$5.00.

particularly in the presence of variations in expertise and workload. Efforts to use decision theoretic techniques to improve MI interactions show promise [12], but much work needs to be done to identify the key components of ideal MI interaction.

When discussing managing teams of robots, it is appropriate to consider the fan-out metric [4]. This metric indicates that autonomy significantly impacts fan-out (via neglect time) by determining how long a robot can operate without human input; similarly, the operator control interface also impacts fan-out (via interaction time) by influencing how quickly the human can instruct one or more robots. Importantly, whether or not a robot can *wait* for human input is a key determinant of fan-out; if the robot can still operate, albeit in a potentially degraded state [5], the number of robots that a single human can manage can dramatically increase [6]. Since MI, DA, and JA control allow for balancing of neglect and interaction times in response to shifting workloads, each control scheme can potentially increase fanout.

Since this paper evaluates large teams of robotics agents, it is appropriate to note that coordination of large-scale teams has received attention [14], particularly in the area of swarm robotics (e.g., [16]) and more recently applied to search-type applications [23]. Additionally, there has been some work on defining levels of team autonomy [9, 22]. However, this area is still new in HRI and we know of no in-depth comparision using large-scale teams of performance between the different types of autonomy and how they're affected by workload or operator expertise. However, it is appropriate to note that WiSAR is similar to urban search and rescue, which has been studied in the context of MI control [2, 17]).

3. EXPERIMENT DESCRIPTION

Participants were given the role of supervisor over a WiSAR mission where the primary task was to find a missing person as quickly as possible. Our short term goal was to study how granting JA, DA, or MI to agents of a large-scale robotic team affects performance. More long term, we wanted to learn when MI works and why. In light of prior work, we evaluated how the level of workload and the participants' prior knowledge affected performance since these factors were previously identified as limitations to the potential benefits of dynamic interaction [11].

Two experiments were performed. Experiments were counterbalanced to avoid learning effects. Subjects were compensated for their time. Participants were given a time limit of 10 minutes for each trial. Each participant completed four trials plus a training mission that explained the purpose of their mission and introduced them to the simulator.

3.1 Apparatus

Each WiSAR trial took place in a simulator. Participants controlled 200 robotic searchers with the goal of finding five missing persons distributed randomly across a map. As illustrated in the simulator screenshot in Figure 1, agents are initially clustered in the center of the interface. 10 backpack items were hidden around each missing person as clues, following a Gaussian distribution. In addition, a number of distracting items were scattered across the map following a uniform distribution. The picture of the missing person is shown on the far right as well as the list of all items from the missing person's backpack. On the left is shown a map of the search area and the distribution of searchers. Between the map and the images from the backpack is a queue of items (backpack items, distractors, images of the missing person) found by the searchers and waiting to be classified by the operator (by pushing the "keep" or "reject" buttons of the user interface).

During experiment one, the number of distracting items was varied to provide different levels of workload. During experiment two, the number of distracting items remained constant, but some participants were given a map with shaded regions, corresponding to areas that had a high probability of containing the missing persons (Figure 2). The shaded regions of the maps simulated imperfect prior knowledge; one of the five clouds was not paired with one of the five missing persons and the distribution of the clouds differed from the actual distribution of the backpack items..

3.2 Descriptions of Autonomy

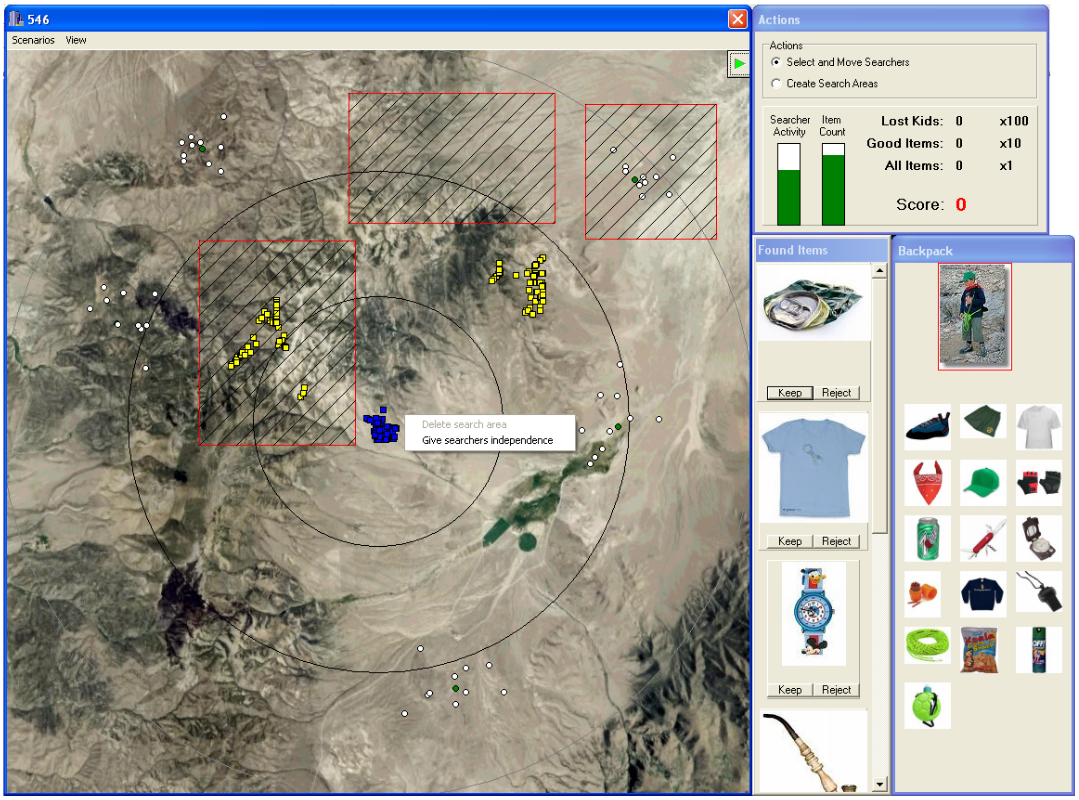
Levels of autonomy. As described in the introduction, an agent's level of autonomy includes the roles that the agent performs. These roles include (a) the amount of responsibility held by the searchers and (b) the amount of authority that the searchers had to change their responsibility. For these experiments, levels of responsibility were high, medium, and low (described below). For DA and MI, searchers had authority to switch between these levels, and for JA and MI the human supervisor had authority to change between these levels. Under DA and MI, we refer to the factors that determine when a switch is made as the *trigger*. Depending on the type of autonomy, participants were able to create and delete search areas, and assign searchers to search areas.

Adaptive autonomy. Under DA, participants had no direct control over the searchers and simply classified items that the searchers found. Searcher agents would begin with high responsibility, performing a 30 meter sweep[2] that focuses on the center of the search area before fanning out to the more distant locations. If the searchers finish sweeping the area without the human classifying an object from the backpack, the searchers selected a new search area and repeated.

The trigger for switching between high responsibility and medium responsibility was when the supervisor identified a backpack item[3].When this happened a subset of searchers would create a smaller, more detailed 6 meter search area centered around the item and conduct a detailed search. At this medium level of autonomy, they would continue to search the area until they found the missing person and the supervisor classified it as such, repeating the search if necessary. The trigger for switching from medium to high responsibility was when the supervisor indicated that the missing person had been found. This level of responsibility is considered medium because the trigger to enter and exit this

[2] When we use the phrase an "N meter sweep", the N refers to the distance between ground searchers. WiSAR operations often include sweeps at various searcher densities [19].

[3] Typical factors that have been used at triggers for switching autonomy in other work include changes in human workload, human physiological signals that indicate stress or negative emotion, robot self-diagnosis that indicates lack of progress, or task-specific signals that indicate performance is lower than expected. Switching autonomy when the human identifies a backpack item is a task-based trigger that mimics changes in operator intent/workload.



Actions
● Select and Move Searchers
○ Create Search Areas

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

Figure 1: The testbed.

level depended on input from the classification task. Under DA, searchers never used the low level of autonomy.

Adjustable autonomy. Under JA, the search supervisor explicitly triggered all changes in the searchers' level of responsibility via a menu. Under high responsibility, the searchers had permission to create their own search areas and perform a 30 meter sweep, successively moving on to new search areas as they completed the previous one. Under low responsibility, searchers would perform a search of an assigned area, awaiting for further instructions when the sweep was completed.

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 medium level. At this medium level, they would continue searching until they discovered the missing person and the supervisor classified it as such, repeating the search if necessary. After the missing person was located and classified, the searchers would return to low responsibility to complete their current search area and find as many backpack items as possible. Although the first experiment allowed the supervisor to cause searchers to perform an 18 meter search, no participants ever used this so it was not included in the second experiment.

Mixed initiative. In MI scenarios, the searchers would start with a high level and return to this level whenever possible. If the search supervisor assigned them to a search area, they would immediately sweep the area; when complete, the searchers would choose their own subsequent search area.

There were two triggers for agents to enter medium responsibility. First, a supervisor could manually create a small (see above) search area around a discovered and classified item. Assigning searchers to it would cause the agents to search until the missing person was discovered and classified. Second, if a discovered item was classified as a backpack item this would trigger the searchers to create their own small search area around it, moving to medium autonomy until the missing person was discovered and classified.

3.3 Metrics

The metrics we used can be broken into two main categories: task performance and management overhead/workload. We assigned a subjective *level of importance* to each metric, relative to their ecological significance as outlined by experts in WiSAR [19] and gleaned from live field trials [8].

3.3.1 Task Performance

For WiSAR, task performance metrics include (a) the primary task of finding and classifying the missing persons, (b) the secondary task of finding backpack items, and (c) supporting tasks such as covering as much ground as possible.

Primary Tasks.

- Average number of missing people classified — high importance. The main measure of success in a wilderness search and rescue domain is whether the missing person is found.
- Probability of success — high importance. This metric, used in the Search and Rescue community, equals the probability that the missing person is contained in a given search area ("probability of area") times the probability that a searcher senses the missing person

given their search paths in the world ("probability of discovery") [8, 19].

We assumed a static distribution for the location of the missing person (probability of area) that peaked directly over the missing person and decayed linearly reaching zero at a radius of 50 cells from the missing person. Although future work should consider other models, this model is a first order approximation to observed distributions from missing person histories [21]. The probability of discovery varied slightly between experiments, but generally followed models developed by WiSAR experts [19]; we will discuss the models when we discuss each experiment.

Secondary Tasks.

- Average number of backpack items classified — medium importance. Participants were told that their primary task was to find all five missing people, but that they should try to find as many of the backpack items as possible.
- Average number of items found — low importance.
- Average searcher distance — low importance. The number of 6m×6m cells moved. For simplicity, searcher speed was constant regardless of terrain.
- Simple coverage — low importance. The combined percentage of the searchable terrain that was covered.

3.3.2 Management Overhead/Workload

Management overhead and workload metrics measure demands on (or performance of) the supervisor. Each of these metrics is derived from a secondary task, either identifying items or managing the interface. Secondary task performance is a way of measuring workload the complements subjective approaches such as NASA-TLX.

- Average number of timesteps to classify items — medium importance. Time is counted between when an item is placed in the classification queue and when the supervisor chooses to keep or reject the item.
- Overhead — medium importance. The amount of time spent on management tasks such as creating search areas, selecting searchers, assigning searchers to search areas, and changing searchers' autonomy level.
- Time to localize — medium importance. The amount of time between between finding the first item in a backpack and finding the missing person.
- Average timesteps inactive — low importance. The number of timesteps when a searcher is completely idle, awaiting instructions from the superviser.

4. EXPERIMENT ONE: HIGH VS. LOW WORKLOAD

The first experiment had a three by two design, comparing performance over the three types of autonomy and two levels of workload (high and low). The high workload condition had 400 distracting items uniformly distributed across the map, and low workload distributed 200 distracting items. 12 people voluntarily participated and performed four trials each, resulting in 48 test cases. For this experiment, the probability of discovery came directly from models developed by WiSAR experts [19]. When searchers performed a 30-meter sweep, they had a 50% probability of finding an object as passed the object, where passing an object included the cell they occupied as well as the two cells to either side. When searchers performed a 6-meter sweep, they had a 90% probability of finding an object in their cell with no probability of finding an object in an adjacent cell. This model preserves relative distances: (6/30 meters matches 1/5 cells).

A two-way analysis of variance was performed. If results for a particular metric are not reported, there is no significant difference across either workload or autonomy. Most importantly, there were no significant two-way interactions between workload and autonomy type suggesting that performance between autonomy types did not depend on the level of workload.

Workload. Of the metrics considered, only the average time to classify items showed a significant difference across the two workload conditions (F(2,31)=5.04, p=0.032). Under JA, classification time was significantly higher for high workload (1726.4 simulator time steps[4]) than low (926.3 time steps) Under MI, classification time between high (2059 time steps) and low workload (511.3 time steps) was even greater. This significant increase indicates that increasing the number of distracting items does indeed increase workload, though the difference in workload was not high enough to prevent the operator from performing well in the primary task metrics.

Differences in Autonomy Types. On the other hand, average searcher distance, simple coverage, and average number of items classified all differed across autonomy types. Incidentally, all three of these metrics are classified as "low importance" in Section 3.3, so we will forgo in-depth discussion and merely say that MI and DA resulted in approximately 21% more distance traveled than JA, and DA resulted in more coverage and more items classified than MI or JA.

Discussion. The differences in the amount of time required to classify items indicates that adding more objects of interest does impact workload, but this change in workload does not significantly impact primary task performance across autonomy types. We discuss this result using the perspective of fan-out.

First, note that for this task performance grows with fan-out; as more agents become involved in the search, the probability that the missing person will be found grows quickly. Second, note that recent work indicates that if progress slows too much while an agent waits for attention from the supervisor, then fan-out can be very small, especially in the presence of variability [6]. Although the secondary tasks had to wait for operator input, there were enough available agents that it was possible to make considerable progress toward the goal even if particular agents had to wait for attention.

Neither the time to find the missing persons nor the probability of success were significantly impacted by an increase in workload. The reason why differs for different types of autonomy. For DA, the agents perform a less intelligent search than they would with human input, but they are continually covering ground, even when the operator has a backlog of objects to classify. As soon as the operator indicates that an object was from the missing person's backpack, which triggers a more refined search, it is likely that other relevant objects or even the missing person has been found. For JA,

[4] Note that time values are reported as time steps so that they can be generalized to varying searcher speeds.

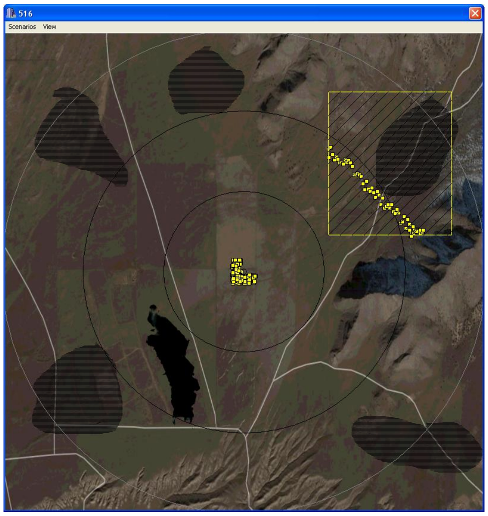


Figure 2: Simulator map with expert clouds.

objects in the classification queue waited longer for operator input than under DA, but the search strategies used by the operators were efficient enough that performance did not significantly degrade.

Thus, we conclude that performance is less susceptible to variations in workload if agents can still make progress toward the objective even when workload causes some tasks to wait for a while for human input.

5. EXPERIMENT TWO: PRIOR KNOWLEDGE

The second experiment also had a three by two design, comparing performance over the three types of autonomy and two levels of prior information (high and low). Low information had no prior information, but high information presented prior information using the shaded areas shown in Figure 2. 18 people participated in the second experiment, resulting in 72 test cases.

Since the first experiment showed that there was little difference in performance between high and low workload, the second experiment adopted a uniformly high level of workload. Additionally, two key parameters were modified between the first and second experiments. First, searcher agents were slowed down. Second, the sensor model for each agent was modified, making it less probable that an agent *would detect* an object; for a 30-meter sweep, the probability of detection was 25% for the searchers cell and the cell on either side, and for the 6-meter sweep the probability of detection was 45% for just the searchers cell. Although these two changes have less claim to ecological validity since the ratios no longer match models from WiSAR experts, the changes make it less likely that the simple relatively coarse search strategy employed when the agents are working under high autonomy will yield success. This makes human input and expertise more important, both in choosing the right responsibility level and in focusing search efforts in the correct areas.

Two-way analysis of variances performed for each metric indicated that with the exceptions discussed below, the following trends hold: (a) none of the metrics showed statistically significant differences between supervisors with and without prior knowledge, (b) all the metrics showed a statistically significant difference across autonomy types, and (c) none of the metrics showed any statistically significant two-way interactions. As a result, we can infer that the prior knowledge of the supervisor did not affect the measured performance, but autonomy management did.

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

5.1 Primary Task Performance

Searchers who employed MI had the highest probability of success at 57.5%and 56.9% for informed and uninformed supervisors, respectivley, although an analysis of variance showed that there was no statistical significance (F(2,62) =1.13, p=0.2919). JA had the second highest probability of success, while DA came in last (Figure 3). This difference across autonomy type was statistically significant (F(2,62)=89.53, p=0.0001).

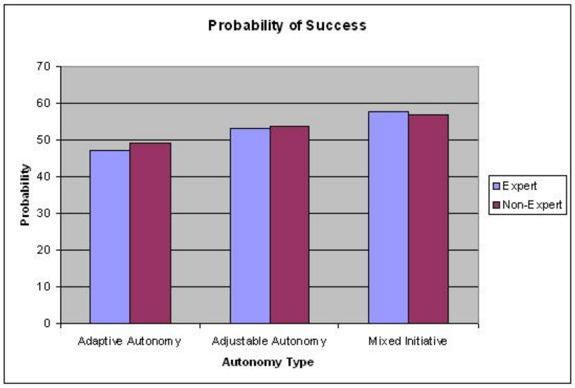


Figure 3: Probability of success

Differences between the number of missing persons found per trial were statistically significant across autonomy types (F(2,62)=7.37, p=0.0014). Figure 4 shows how JA and MI resulted in the most missing persons found on average, while DA did significantly worse. This poor performance by DA can be directly related to the searchers' lack of prior knowledge and poor decision making in their choice of which areas to search.

5.2 Secondary Task Performance

Giving the searchers DA allowed the supervisor to classify the most items (Figure 5). This is statistically significant (F(2,62)=28.0, p=0.0001). However, this metric does not tell the whole story. 88% (400 out of 455) of the items were distracting items. DA'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 MI and JA, coverage was less inclusive but focused on areas with high probability of containing the missing person and their backpack items. Such focused coverage

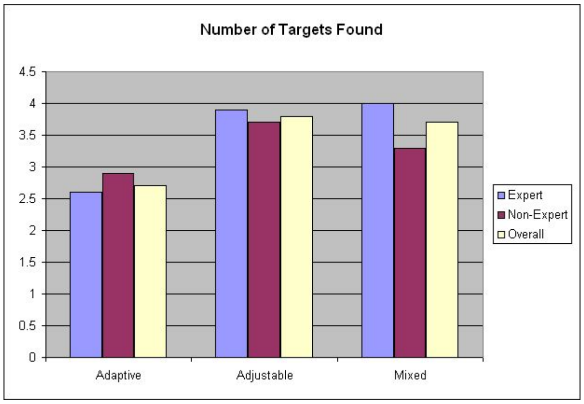


Figure 4: Total number of missing persons found

occurred both because of prior search information (e.g., the search clouds) and because of the more effective response to a known object of interest (e.g., focusing more agents on an object from a backpack). This enabled MI and JA to outperform DA when just backpack items were considered (Figure 6). This is also statistically significant (F(2,62)=4.55, p=0.0143).

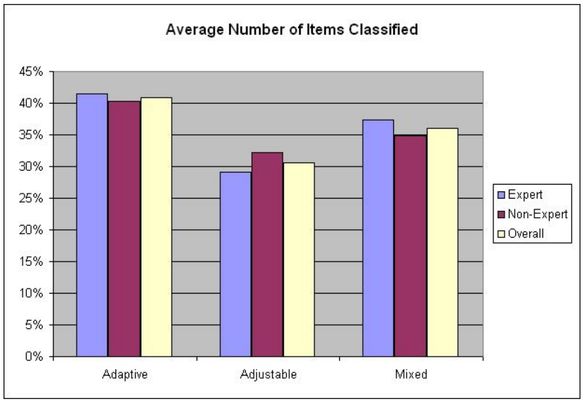


Figure 5: Average percent of items classified

DA and MI resulted in the most distance covered by the searchers (Figure 7). 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. JA autonomy performed on average 15% worse in this metric than DA. When searchers 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(2,62)=54.12, p=0.0001).

Differences in coverage were statistically significant (F(2,62) =52.61, p=0.0001). DA resulted in almost 90%, followed by 63% for MI and below 55% on average for JA. In addition, the differences between informed and uninformed supervisors were statistically significant (F(2,62)=7.88, p=0.0067) for JA (t=12.18, p=0.05) and MI (t=3.08, p=0.05). Under both these types of autonomy, informed supervisors would focus their search on the search clouds, significantly reducing the amount of area they covered though keeping their travel distance high.

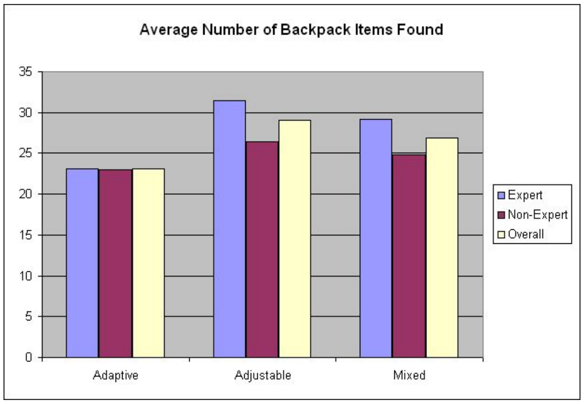


Figure 6: Average number of backpack items found

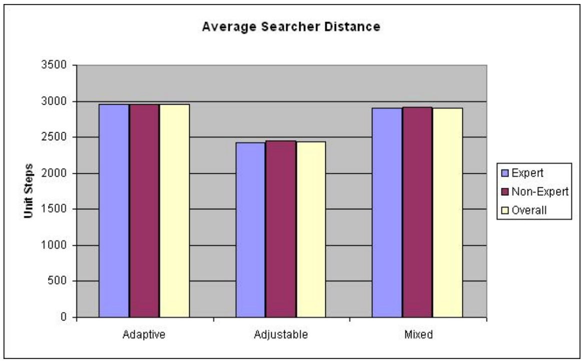


Figure 7: Average searcher distance

Note that coverage is not necessarily proportional to searcher distance. Since the searchers operating under MI traveled as much distance as the searchers using DA, the difference in the amount of coverage is due to the searchers operating under MI repeatedly covering the same ground. This is not necessarily bad in practice, since a single pass over a given search area does not result in a 100% probability of detection and some areas have a higher probability of containing the missing person than others.

5.3 Management Overhead/Workload

Searchers operating under DA and MI are never inactive by design, but searchers under JA sat idle 18% of the time. This was due to high operator workload (the supervisor unable or choosing not to respond), a lack of supervisor awareness when searchers finished a task (the supervisor unaware of a need to respond), and difficulty in finding the inactive workers because of interface design or other reasons.

DA served as a baseline measurement for classification time, since the supervisors had no other workload than classifying items. Speed was correspondingly high, on average 163 timesteps. When searchers were using JA, supervisors had to give them a lot of attention and their performance suffered (923 timesteps per classification). Using MI, 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 8). These results were statistically significant (F(2,62)=5.97, p=0.0043).

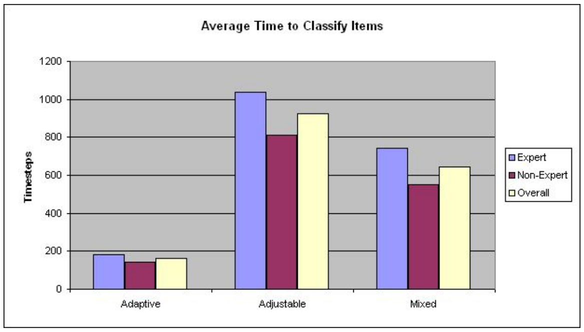


Figure 8: Average time to classify items

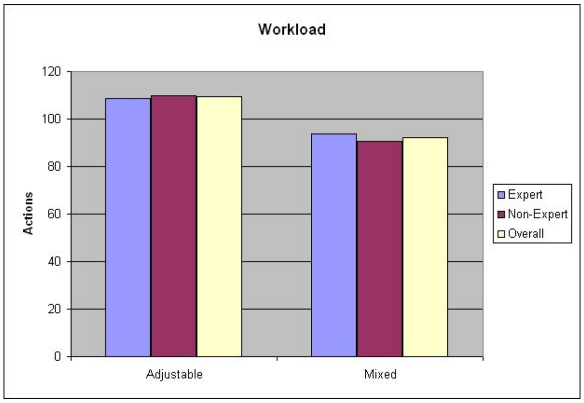


Figure 9: Workload

Figure 9 shows the supervisor's management workload including selecting 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. Only results for JA and MI are reported, since DA supervisors were not given the ability to perform any of the aforementioned actions. This difference in workload is statistically significant (F(2,62)=54.75, p=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.

5.4 Discussion

Table 1 summarizes the results from all individual metrics, including those that were not discussed in detail. In general, MI resulted in better performance than scenarios where adaptive or JA were used. Although MI did not rank first in several metrics, it ranked first in more of the metrics than either DA or JA, 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 metrics — often a very close second. We conclude that MI was better than DA and JA for this experiment.

For generalization purposes, it is useful to discuss why MI outperforms JA and DA in WiSAR given the specific autonomy implementations, since prior work indicates that this is not always the case. The first reason is that there were complementary abilities between the supervisor and the autonomous agents. This is apparent by noting that JA and MI were strongest in metrics falling under the primary and secondary task performance categories, while DA fared especially well in management overhead metrics. This observation is reinforced by noting that most human supervisors chose to complete their primary task of finding all five missing people before focusing on the secondary task of finding as many backpack items as possible. Qualitative observations indicate that while supervisors focus on localizing a single person, autonomous agents continue to cover ground in other areas, striking a good balance between coverage and detection.

Table 1: Overall Autonomy Rankings

High Importance	DA	JA	MI
Prob of success		2nd	1st
Avg # missing people classified		1st	1st
Medium Importance	DA	JA	MI
Avg # backpack items classified		1st	2nd
Avg time from item to owner	1st		2nd
Avg# of timesteps to classify items	1st		2nd
Var between terrain types		2nd	1st
Workload	N/A	2nd	1st
Low Importance	DA	JA	MI
Avg searcher distance	1st		1st
Simple coverage	1st		2nd
Avg timesteps inactive	1st		1st
Avg # items classified	1st		2nd

The second reason that MI outperforms JA and DA is that the problem of micro-management, identified in prior work [5], did not significantly decrease performance. Recall that the supervisor's level of prior knowledge made little difference in the results; informed and uninformed supervisors were helped or hindered by the type of autonomy equally in most metrics. When searchers were operating under MI, the supervisor would often override the searcher's choice of actions with one they deemed more effective. This often resulted in a waste of the searcher's time and travel, but if the supervisor possessed prior knowledge, this overriding resulted in a better end-performance. However, qualitative observations suggest that supervisors often failed to take travel time into account, often assigning searchers to search areas a long distance away rather than assigning uncommitted searchers based on their proximity to the desired area. Although better training or more experience might reduce this negative effect, the effect suggests prior knowledge about what areas to search is counterbalanced by a poor assignment of agents to tasks. Thus, the complementary abilities of search agents and the supervisor tend to cancel problems associated with micro-management.

The third reason that MI outperforms JA and DA is that progress can be made without agents having to wait for extensive periods of time for human input. Simply put, there were enough agents that even if some were being closely controlled by a human, other agents could use environmental triggers to manage their roles in a way that allowed them to contribute to the search. Moreover, the interface made it relatively easy to manage agents; interaction times were low since the supervisor could easily understand agent intent and easily task agents.

6. CONCLUSIONS

Theoretically, agents operating under mixed initiative have all the initiative of adaptive autonomy, while still giving the supervisor the control flexibility of adjustable autonomy. Provided that agents and the supervisor have complementary abilities, this can potentially make mixed initiative (MI) control superior to adaptive autonomy (DA) and adjustable autonomy (JA). For this experiment, MI outperforms JA and DA suggesting that it is possible to achieve the theoretical benefits of MI control for some problem domains, even though some prior work indicates that problems of MI can sometimes outweigh its benefits.

Two aspects of these specific experiments allow MI to outperform DA and JA. First, the performance on the task was proportional to fan-out, high-fan-out requires consistent progress by agents, and MI allowed agents to make progress without waiting too long for human input. Second, the nature of the task and the design of the interface allow the operator to easily perceive agent progress and intent, and to easily task multiple agents to perform the task.

The results of these particular experiments suggest more general lessons for applying MI to large-scale robotic teams. In particular, they indicate that MI can approach its theoretical benefits but that at least three requirements must be met. First, agents and the supervisor must have complementary abilities. Second, agents must be able to progress without waiting for human input. Third, the human must be able to interact efficiently with large numbers of agents. Future work should revise these lessons and evaluate how generally they can be applied.

7. ACKNOWLEDGMENTS

This work was partially supported by the National Science Foundation under grant number 0534736 and by a grant from the Army Research Laboratory. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the sponsoring organizations.

8. REFERENCES

[1] 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.

[2] D. Bruemmer, R. Boring, D. Few, J. Marble, and M. Walton. 'I call shotgun!': An evaluation of mixed-initiative control for novice users of a search and rescue robot. In *Systems, Man and Cybernetics*, volume 3, pages 2847–2852, 2004.

[3] D. Bruemmer, D. Dudenhoeffer, M. Anderson, and M. McKay. A robotic swarm for spill finding and perimeter formation. In *Spectrum*, 2002.

[4] J. W. Crandall, M. A. Goodrich, D. R. O. Jr., and C. W. Nielsen. Validating human-robot interaction schemes in multi-tasking environments. *IEEE Transactions on Systems, Man and Cybernetics — Part A: Systems and Humans*, 35(4), 2005.

[5] M. L. Cummings, C. E. Nehme, J. Crandall, and P. Mitchell. *Developing Operator Capacity Estimates for Supervisory Control of Autonomous Vehicles*, volume 70 of *Studies in Computational Intelligence*, pages 11–37. Springer, 2007.

[6] M. A. Goodrich. On maximizing fan-out: Towards controlling multiple unmanned vehicles. In M. A. Barnes and A. W. E. III, editors, *Human-Robot Interaction in the Military*, page To appear. 2008.

[7] M. A. Goodrich, D. O. 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.

[8] M. A. Goodrich, B. S. Morse, D. Gerhardt, J. L. Cooper, M. Quigley, J. A. Adams, and C. Humphrey. Supporting wilderness search and rescue using a camera-equipped mini uav: Research articles. *J. Field Robot.*, 25(1-2):89–110, 2008.

[9] J. Hackman. Humans, robots, and teams, 2007.

[10] 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.

[11] 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.

[12] T. Kaupp and A. Makarenko. Decision-theoretic human-robot communication. In *Proceedings of HRI2008*, 2008.

[13] D. Kortenkamp, R. Bonasso, D. Ryan, and D. Schreckenghost. Traded control with autonomous robots as mixed initiative interaction. In *Mixed Initiative Interaction*, 1997.

[14] A. A. Makarenko, T. Kaup, and H. F. Durrant-Whyte. Scalable human-robot interactions in active sensor networks. *Pervasive Computing*, pages 63–71, 2003.

[15] C. Miller and R. Parasuraman. Beyond levels of automation: An architecture for more flexible human-automation collaboration. In *Human Factors and Ergonomics*, pages 182–186, 2003.

[16] 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.

[17] R. Murphy, J. Casper, M. Micire, and J. Hyams. Mixed initiative control of multiple heterogeneous robots for urban search and rescue. In *Robotics and Automation*, 2000.

[18] 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*, pages 208–213, 1999.

[19] T. Setnicka and K. Andrasko. *Wilderness search and rescue*. Appalachian Mountain Club, 1980.

[20] T. Sheridan and W. Verplank. Human and computer control of undersea teleoperators. In *MIT Man-Machine Laboratory*, 1978.

[21] W. G. Syrotuck. *An Introduction to Land Search: Probabilities and Calculations*. Barkleigh Productions, Mechanicsburg, PA, 2000.

[22] J. Wang and M. Lewis. Human control for cooperating robot teams. In *ACM SIGCHI/SIGART Human-Robot Interaction*, pages 9–16, 2007.

[23] E.-M. Wong, F. Bourgault, and T. Furukawa. Multi-vehicle Bayesian search for multiple lost targets. In *Proceedings of the 2005 IEEE International Conference on Robotics and Automation*, pages 3169–3174, Barcelona, Spain, April 2005.