Everybody Needs Somebody Sometimes: Validation of Adaptive Recovery in Robotic Space Operations

Steve McGuire[1], P. Michael Furlong[2], Terry Fong[3], Christoffer Heckman[4],
Daniel Szafir[4], Simon J. Julier[5], and Nisar Ahmed[1]

***Abstract*—This work assesses an adaptive approach to fault recovery in autonomous robotic space operations, which uses indicators of opportunity, such as physiological state measurements and observations of past human assistant performance, to inform future selections. We validated our reinforcement learning approach using data we collected from humans executing simulated mission scenarios. We present a method of structuring human-factors experiments that permits collection of relevant indicator of opportunity and assigned assistance task performance data, as well as evaluation of our adaptive approach, without requiring large numbers of test subjects. Application of our reinforcement learning algorithm to our experimental data shows that our adaptive assistant selection approach can achieve lower cumulative regret compared to existing non-adaptive baseline approaches when using real human data. Our work has applications beyond space robotics to any application where autonomy failures may occur that require external intervention.**

***Index Terms*—Human-Centered Robotics, Space Robotics and Automation, Learning and Adaptive Systems**

I. Introduction

FUTURE space missions will feature autonomous robotic and human crewmates. However, research in autonomous systems has yet to produce a robot that never fails. Therefore, in many planned space missions, robots must be able to request assistance from other robots and humans. This work examines the open problem of how robots choose an assistant to recover from failures thereby maximizing the amount of time spent operating autonomously. In our previous work [1], we investigated how a robot may optimize its choice of assistance through the use of *indicators of opportunity* (IOO). IOOs are predictors of the performance of each potentially available assistant. IOOs could, for instance, include physiological state measurements (e.g. blood pressure, heart rate, galvanic skin response, O_2 levels, etc.) as well as contextual indicators (e.g. location, current task assignment and workload, past performance on assigned tasks). IOOs are created and constantly revised using data which are already being collected for some other reason. The use of IOOs for informing assistant selection is preferable to the use of dedicated sensors in order to minimize impact on mission design, cost, and execution.

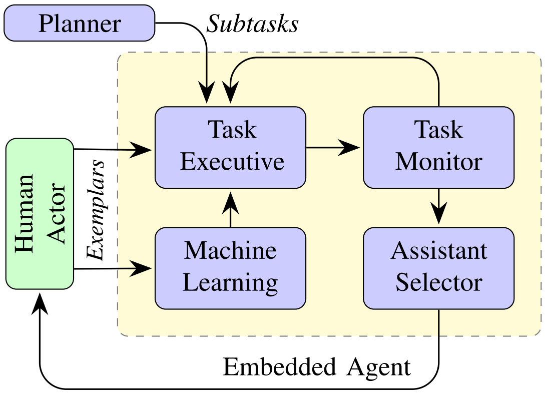


Fig. 1: *Information flow in an idealized learning robot. The planner decomposes goals into subtasks for the executive.*

In [1], we validated the use of IOOs in an idealized simulation of an autonomous robot working together with human assistants in a space environments, using a context-aware reinforcement-learning (RL)-based policy for adaptive assistant selection. Using cumulative regret (the running sum of differences between rewards earned by a policy's choice of assistant and the highest possible rewards that could have been earned) as a measurement of performance, we demonstrated that the IOOs can be exploited with "black box" models of human performance to provide improved performance, as compared to existing baseline assistant selection policies that do not consider such contextual information.

In this paper, we extend our earlier work by applying and evaluating our RL-based approach to a real human IOO and performance data set for a range of failure recovery tasks that are representative of future space robotics mission scenarios. This data set was acquired through a set of human subject experiments that simulate the operating conditions that humans are expected to operate in, capturing several important features for assessing the feasibility of an adaptive policy selection approach for future mission designs.

Our paper has three major contributions. First, due to the

Manuscript received: September 10, 2018; Revised December 7, 2018; Accepted January 3, 2019.

This paper was recommended for publication by Editor Allison Okamura upon evaluation of the Associate Editor and Reviewers' comments. This work was supported by a NASA Space Technology Research Fellowship.

[1] Steve McGuire and Nisar Ahmed are with the Department of Aerospace Engineering Sciences, University of Colorado at Boulder, USA stephen.mcguire@colorado.edu

[2] P. Michael Furlong is with SGT, Inc., NASA Ames Research Center

[3] Terry Fong is with NASA Ames Research Center.

[4] Christoffer Heckman and Daniel Szafir are with the Department of Computer Science, University of Colorado at Boulder.

[5] Simon Julier is with the Department of Computer Science, University College London

Digital Object Identifier (DOI): see top of this page.

expense and difficulty in sourcing human performance data, we present a technique to synthesize IOO and task performance training/testing data from a small number of samples. We devised and executed an experiment to statistically validate our synthesis technique, finding no evidence to discourage the use of our method. Second, since we have no previously validated data set to use for our assistant allocation problem, we present our own design and validate using human subjects that there are useful relationships between IOOs and human performance to be learned . Finally, we test the proposition that our technique can learn these relationships. We find that our learning-based approach can exploit the embedded information to yield improved performance over existing non-adaptive assistant selection policies that have been proposed for failure recovery in autonomous robotic space operations.

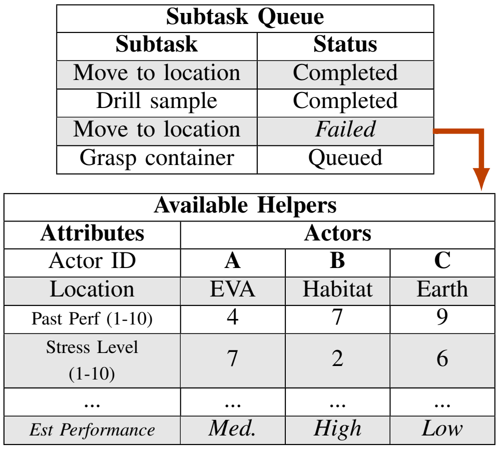
II. Background and Problem Statement

This work is motivated by a hypothetical Mars mission scenario derived from existing Earth studies [2] and proposed future manned exploration missions [3]. The crew has been given a set of high-level *goals* which define the mission. Each goal is comprised of a set of *tasks* that must be completed. These are decomposed into a series of *subtasks*, atomic units of work which cannot be subdivided further [4]. For example, the subtasks for a mobile rover obtaining a sample from to an area of interest might include: moving to the sample location, drilling to obtain a sample, moving a sample container to an onboard storage area, and then moving to the home habitat. The inability to complete a subtask to specification is called a *failure*. It is recognized by the *task monitor*, a part of the onboard embedded agent. [5], [6], [7]. For example, Fig. 2 illustrates there was a failure during a “Move to location” subtask.

In the event of failure, the *assistant selector* must chose an appropriate *assistant* to address the failure and resume autonomous operations according to the rules of a *policy*. Assistants may be consulted for two reasons: task recovery and task demonstration. In this paper we focus solely on task recovery. A key constraint on this recovery process is to recognize that there is a cost associated with using humans as sources of aid, as they will have other primary responsibilities besides monitoring robotic operations.

The current practice in space robotic operations is to dedicate human operators to monitor and assist all the robots all the time [8], [9]. These individuals are chosen through a combination of expert opinion and design choice. One common approach is to use a *static policy*. Given a static allocation table (where an offline analysis has considered the cost and benefit of assigning each task to each human [10]), an optimal assignment is carried out. However, this static approach cannot meet the dynamic demands of a real situation, in which an operator’s attention becomes divided between multiple competing robots or external tasks [11].

An alternative is to use dynamic assistant allocation. Rather than use an assignment based on fixed costs and benefits, these mechanisms constantly revise the costing and update the allocation strategy in response to evolving human capabilities and constraints. As shown in Fig. 2, robots can choose from a number of possible assistants whose quality is each a function of dynamic features. Robots must learn to select the best assistant given their current state.



Subtask Queue	
Subtask	**Status**
Move to location	Completed
Drill sample	Completed
Move to location	*Failed*
Grasp container	Queued

Available Helpers			
Attributes	**Actors**		
Actor ID	**A**	**B**	**C**
Location	EVA	Habitat	Earth
Past Perf (1-10)	4	7	9
Stress Level (1-10)	7	2	6
...	...	...	...
Est Performance	*Med.*	*High*	*Low*

Fig. 2: *Robot obtains assistance with ‘move to location’ subtask. Robots can query a number of helpers with different attributes, like experience level, cognitive/physical workload, location constraints, and the cost for disrupting humans.*

A. Formal Assistance Allocation Problem Statement

Assume there exists a set $\mathcal{A} = \{a_v\}_{v=1}^{N_\mathcal{A}}$ of actors capable of assisting an embedded agent during an autonomy failure. The embedded agent a_u maintains a history from timestep $k = 1$ to the present of past subtasks t_k that have failed, assistant allocation decisions, and resulting reward earned by each decision.

At timestep k, a_u must assign a single actor a_j from set $\bar{\mathcal{A}} = \mathcal{A} \setminus \{a_u\}$ of size $N_{\bar{\mathcal{A}}}$ to recover from failed subtask t_k and maximize mission utility U, given only the partially observed history from timesteps $1, ..., (k-1)$ of past assignment decisions and resulting rewards. That is, maximize $U = \sum_{k=1}^{\infty} \gamma_k \mathbf{r}_k$, $\gamma_k \in \{0,1\}^{N_{\bar{\mathcal{A}}}}, |\gamma_k| = 1$ where γ_k is a vector of indicator values denoting the selection decision for each timestep k and $\mathbf{r}_k$ is the reward vector earned by every agent at timestep k, subject to the subject to the constraint $\gamma_k \cdot r_k = |r_k|$, meaning only the selected performance is observed at each timestep k. The rules for composing γ_k describe a *policy*. Several simplifications and restrictions are present in our problem. No subtask may be assigned to more than one actor, nor are embedded agents allowed to help one another. Furthermore, subtask failure distributions are unknown, and can’t be used in actor selection.

B. Related Work

Classical centralized [12], [13] and distributed [14] task allocation schemes require coherent utility functions to model the overall expected payoff of actor-task assignments; developing such utility functions requires comprehensive domain knowledge that is unavailable in a space exploration setting due to operational uncertainty. Importantly, most allocation schemes assume that the underlying task execution process and associated utilities are well-modeled and time-invariant, making them brittle when the underlying system is dynamic or only partially understood [15].

To attempt to overcome this brittleness, some authors have used partially observable Markov decision processes (POMDPs) to learn offline a policy which can be used online to determine an optimal actor assignment policy. POMDPs have also been used to optimize spatio-temporal assignments of robots to tasks [16] and other human-robot collaboration efforts [17]. However, as above, POMDPs require accurate system models to evaluate both actions and rewards. Optimal allocation policies are also extremely difficult to find for POMDPs with high-dimensional state spaces [18].

In contrast to model-based approaches, Reinforcement learning (RL) can adaptively integrate feedback from action choices to improve future choices. Parker, et al's behavior-based L-ALLIANCE [19] is an early example of an RL-based task allocation system. However, L-ALLIANCE only considered time to complete objectives, when in fact time is only one of many factors that influence overall task and mission performance. Nevertheless, RL algorithms allow contextual and performance data from assigned actors to be leveraged over time, without requiring accurate *a priori* knowledge of system models, making RL-based techniques attractive for exploration missions with highly uncertain models.

The deployment of autonomy techniques in actual missions has been rather limited; one shining success story is the Automated Exploration for Gathering Increased Science (AEGIS) system aboard NASA's *Curiosity* Mars rover [20]. The AEGIS system autonomously selects science targets to be analyzed via laser spectrometer without requiring a round-trip to Earth; a clear benefit of this system is increased science output due solely to the use of autonomy. While the type of autonomy implementing AEGIS is comparatively basic due to limited onboard compute, the acceptance of autonomy in science operations represents a major cultural shift and bodes well for future deployment of learning techniques such as ours.

Our previous work used a hybrid linear contextual multi-arm bandit [21], using the *indicators of opportunity* as *context features* to inform assistant selections. The hybrid formulation lets us consider observations that affect all actors, as well as individual observations capturing unique human responses. We validated the utility of contextual information in making decisions by comparing a non-parametric multi-arm bandit approach (KLempUCB, [22]) with a context-aware approach, finding a marked performance improvement in terms of cumulative regret. However, our previous work used a synthetic human performance model that was not grounded in real-life data; in this work, we demonstrate the effectiveness on actual human subjects.

III. Experimental Designs and Results

Accurately assessing our proposed algorithm requires data from the behaviour of human assistants. We have collected a dataset of human subject performance to fill a gap in available datasets. As in our prior work, we measure performance as *cumulative regret*. Regret is the difference in reward between the optimal choice (unknown to the policy) and the policy's choice of assistant. Cumulative regret is the sum of the regret at each time step. The units of the reward function, and hence regret, depend on the system design. Our adaptive policy's performance is compared to the state-of-the-art static approach of [10] through the *crossover* point, which is the number of selection events required for the cumulative regret of the static policy to exceed the cumulative regret of the adaptive policy. This metric allows us to measure the performance penalty incurred by the exploration of the learning algorithm

A. *Synthesizing large datasets from a small n*

One of the major challenges of this work was forming assistant pools (sets of potential assistants) so regret could be used as a metric. Assistant selection algorithms should be tested against a number of different subjects to ensure generalizability to other users. In our previously published work, we used simulated actors. However, these do not accurately model the behaviour of real humans. Practically, human testing is very limited in the number of subjects that can be enrolled. Matching a baseline of one hundred trials with three actors would require three hundred subjects. To address a similar problem, [21] samples possible actions uniformly online; offline, their algorithm skips data samples until the desired action is taken. Such an approach is impractical because the number of data samples needed scales linearly in the number of available actions to obtain a given set of desired observations. Other researchers address the small n problem by using techniques that are unavailable in our problem, such as the use of reflection and rotation in the object recognition domain.

Instead, we use a technique from the field of resource utilization [23] to select smaller subsets of fixed size from a population in order of increasing overlap between subsets. This technique balances subset membership uniformly so as to ensure that our entire test population is well-represented and to avoid 'cherry-picking' the actor pool. We rely on an assumption that each actor's performance is independent of inter-actor effects. We implemented the minimal-overlapping algorithm from Section 3.3 of [23], which exhaustively computes the number of common members between the set of already selected teams (actor subsets) and every potential next member, and computes an index of overlap for candidates. The next candidate member is chosen from the candidates with minimum index. This minimal-overlapping algorithm was chosen for its intuitiveness and simplicity. We want to show that artificially increasing the number of trials does not negatively affect the ability to assess algorithm performance by introducing biases. We therefore carried out an experiment to ensure the validity of our technique.

Hypothesis 1: *The use of teams with overlapping members impacts the performance of assistant selection algorithms.* If we reject the null hypothesis, then our technique for synthesizing additional data results in a statistically significant change in the distribution of crossover points; such a change would indicate that our technique is not suitable for use. However, if we fail to reject the null hypothesis, we can gather evidence to indicate that our technique is safe for use even though we cannot prove the null hypothesis.

We conducted a 2×1 experiment to determine if using shared teams affected the distribution of crossover points compared to independent teams. The independent variable was the choice of independent simulations for each actor in each team or an enumeration of maximally disjoint simulations in which teams were composed of shared simulations of actor performance. Our dependent variable is the crossover point of our adaptive policy relative to the static policy.

Two matched sets of simulations were conducted. In each set, fully disjoint teams of actors were compared against shared teams selected from a set of ten actors, producing regret curves similar to our previously published work [1]. The number of selection events needed before the multi-arm bandit outperformed the informed static policy was then measured for one hundred trials in each setup. This procedure was repeated for teams of three and six actors.

a) Results: We used a Kruskal-Wallis nonparametric test [24] to determine if the distributions of crossover points over selection events between the two conditions was significant in both the three-actor case and the six-actor case. In one hundred trials, there was no difference between the distributions for either the use of three actors (H(79) = 80.515, with p = 0.43) or for the use of six actors (H(72) = 74.574, with p=0.39). These densities are shown in Figs. 7 & 8 in the Supplementary Materials. These tests were only run in the trials when crossover occurred, omitting 14 trials in the 6-actor case which the multi-arm bandit never underperformed the informed static policy. Since we cannot reject the null hypothesis that the distributions are identical, we fail to accept Hypothesis 1, concluding that the distribution of crossover points are similar enough to permit the use of the alternative test set composition of the maximally disjoint subsets. Having reached this conclusion in simulation, we may then apply the result to a study with real humans. While this conclusion includes the possibility of committing a Type II error (false negative), we have observed no evidence to this effect based on visual inspection. A stronger hypothesis test would have been preferable; however, the crossover points do not appear to follow a well-known distribution.

One worthwhile observation is that the generation of the one hundred most disjoint teams from space $\binom{10}{3}$ required two shared members (e.g. teams A,B,C and A,B,D were present) and in the space $\binom{10}{6}$ required four shared members (teams A,B,C,D,E,F and A,B,C,D,G,H). Even though the teams share more than half their members, the crossover point distributions are still not statistically different.

B. Development of a reference human performance dataset

In previous work [1], we used idealized simulations to show that indicators of opportunity can improve the performance of an assistant selection algorithm. Our intent in developing a human-sourced dataset is to capture the difficult-to-model properties of human task performance. This reference dataset lets us make a fair comparison between strategies in a manner that mirrors the technique used in the proof-of-concept simulations, except that we can now include effects such as workload, attention, and fatigue.

In our experiment, humans assist virtual robots in a series of tasks relevant to planetary exploration; each human subject is scored on the basis of their performance for each task. This dataset captures individuals' performance responses to both explicit manipulations (Fig. 3), including workload and attention, and implicit manipulations, such as task difficulty. Performance is measured by scoring results of each planetary exploration-relevant tasks by our software. Several research questions are proposed to explore the feasibility of using of an adaptive allocation strategy with real humans and develop heuristics for implementing a real system.

Question 1a: *Are our experimental manipulations effective in manipulating human task performance?* We expect to see a significant effect on human performance as a result of our experimental controls.

Question 1b: *What physiological responses, such as galvanic skin response and heart rate variability, can be used to predict human performance?* This question explores the effectiveness of directly inferring (in contrast to the indirect estimates of [25]) performance from two physiological measures with established relationships to task performance: heart rate variability and galvanic skin response. These measurements are already being collected in current space operations and thus are indicators of opportunity, adding no additional mission design burden. Both of these measures have shown promise for individuals on a given task, but are inconclusive when generalized across a population and multiple tasks. Many other data types are also available - we fully expect to find richly interconnected dependencies between data and human performance. This question explores whether these measures are effective in predicting individualized performance on a task-by-task basis.

Question 1c: *How important are individual responses when attempting to model human performance?* This question examines whether our experimental design captures sufficient information to enable algorithms that consider individual differences to outperform those that do not. This is, for instance, motivated by Cacioppo [26], who offers a model of psychophysiology partitioning observed physical traits that inform cognitive state into sets that generalize across all humans and those that are particular to specific humans.

Our experiment is structured as a mixed design; within-subjects independent variables are workload and attention task parameters, while between-subjects independent variables are task type and task difficulty. A mixed-design experiment is appropriate in our case because it can achieve large power with a small numbers of subjects. The transfer learning effects common to mixed-design experiments are expected: they are important if we seek to capture the richness of human learning and task performance.

Our between-subjects variables are necessary to present the illusion to a selection policy that every selection trial is conducted under the same conditions for every possible actor; without this illusion, our dataset would not be usable for policy development. Workload, attention, and task difficulty are each structured as having three levels, while we include three task types: *navigation*, *sample handling*, and *grasping*. The dependent variable is *productivity*, computed as the ratio of performance score and time required, while the controls consist of two within factors: workload and attention, and two between factors: task type and task difficulty. Our design is summarized in Fig. 3. To capture individual physiological responses, heart rate variability and galvanic skin response data were continuously measured over the course of the experimental protocol using a Microsoft Band 2 wireless monitor.

a) Experimental Tasks: This experiment includes three planetary exploration mission scenarios where a user recovers a failed autonomous robot while attending to other primary attention tasks. Each assistance task has a continuous quality metric, rather than a binary "success/failure" outcome. Also, each component task supports three difficulty levels. Each task defines a quality score $\mathbf{q}$ as well as a resource cost $\mathbf{c}$. Our reward metric, *productivity*, is calculated as the ratio of the earned score to the resources expended to earn that score.

In the *Navigation* assistance scenario, the user teleoperates a robot across a terrain to a goal, shown in Fig. 4a. This scenario is driven by a Gazebo physics simulation of a exploration rover exploring an environment, sourced from Mars Express data. The user is presented with an overhead view including current position, past positions, and desired goal, as well as a first-person view from a simulated on-board camera system. The user's performance ($\mathbf{q}$) is quantified as the fraction of the distance remaining to the goal to the Euclidean distance between the starting position and desired goal position, while cost ($\mathbf{c}$) is the total time to execute.

In the *Grasping* scenario, the user aids the robot in grasping differently shaped objects using a parallel gripper. This scenario is a simple approximation of a grasping task, shown in Fig. 4b. Grasp quality $\mathbf{q}$ is determined by the alignment of the parallel jaws to object facets. The cost, $\mathbf{c}$, is the elapsed time to complete the grasp. Subjects use keyboard arrow keys to control end effector position and rotation; difficulty is varied by changing object geometry and pose.

In the *Sample Handling* scenario, the user aids the robot in depositing recovered solid sample material into the funnel of an on-board scientific instrument in the presence of wind. The goal is to transfer all material recovered by a drilling operation into the analysis funnel. This scenario is driven by a first-order approximation of particle motion in wind as a 2D Gaussian distribution, shown in Fig. 4c. The user is presented with an overhead view of the funnel, with an x-y marker in the plane to denote the current center of mass of the scoop containing sample material and a wind observation. The user teleoperates the sample delivery scoop into position relative to the center of the funnel. A simulation then determines the percentage of material that was recovered by the funnel. Difficulty is controlled by the wind parameters: velocity, gust behavior, and observation rate. Quality $\mathbf{q}$ of the user's task performance is the percentage of recovered material, while the cost $\mathbf{c}$ is the elapsed time used to position the scoop.

While we expected to observe changes in performance due solely to the difficulty of the subtasks, additional independent controls on user attention and workload directly manipulated cognitive state via distractor tasks, impacting performance. The assistant selection algorithm used these control settings in making decisions. To manipulate attention, subjects were presented with messages to be classified as high or low priority. To manipulate workload, subjects performed a variation of the NASA workload simulation tool Multi-Attribute Task Battery II (MATB-II) [27], where a variable liquid tank level must be kept within bounds by operating multiple pumps. Exponential distributions with rates λ control the time between distractor tasks. The conditions were manipulated via the parameters λ of the distributions; subjects received a randomized ordering of several fixed values of λ. The values of λ were fixed for a constant number of subtasks. To establish a baseline for task performance, we included an initial training period with no distractor tasks.

b) Apparatus and Environment: A key challenge for this experimental design was ensuring that subjects completed a sufficient number of subtasks, so that our assistant selection algorithm had enough data to learn from. Based on our prior simulations [1], we found that the number of selection events required to achieve parity with a state-of-the-art approach scales linearly in both the size of the assistant population and the number of types of subtask. Therefore, the human data collection involved a small number of subtasks and used small assistant populations, to validate performance of our adaptive approach within the horizon imposed by the number of selection events possible in a single testing session.

A list of 150 specific subtasks was generated by sampling subtask types uniformly. The difficulty of each subtask was sampled uniformly. This list of scenarios was then presented identically to each subject. Workload and attention variations were determined individually per subject. The list of subtasks were broken into 30 blocks of 5 subtasks. For each block, two parameters were sampled from $Uniform(0, 0.1) : \lambda_a$ and λ_w. Within each block, distributions $Exponential(\lambda_a)$ and $Exponential(\lambda_w)$ were sampled to obtain time between attention and workload distractor tasks respectively. The first block was hard-coded to have no workload or attention tasks to provide a training period. Fig. 3 summarizes our controls, human-sourced data, and observed measures.

c) Subjects: A total of 17 subjects were recruited from the University of Colorado community. Our subject pool included both technical and non-technical users, all of whom prefer right-handed mousing controls. After obtaining consent, users were briefed on the experimental tasks and user interface, fitted with their fitness tracker, and seated in a quiet office-like environment in groups of up to four subjects. Each subject was assigned an individual workload-attention profile and then left to complete the task sequence at their own pace, taking approximately three hours to complete. Once complete, each subject provided feedback, was compensated, and then dismissed. Our collection methods were approved by IRB under protocol 17-0485.

d) Results: This experiment is structured as a hierarchical linear model, where a linear mixed-effects parametric model with random slopes is used to model productivity on a per-subject and per-task basis. We used this model type to account for the fact that our productivity data are grouped by task type and task difficulty. These statistical analyses provide a 'best case' ground truth of the predictive capability of our controls and physiological measurements, and ensure that our dataset encodes a relationship between observations and productivity that can be learned.

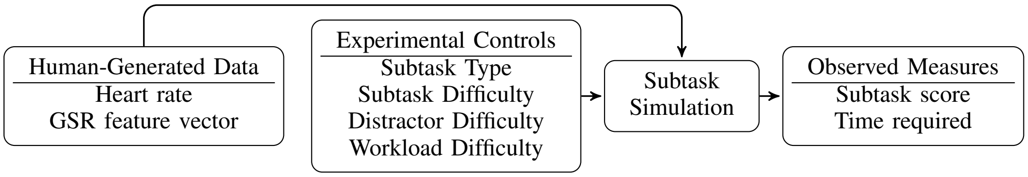
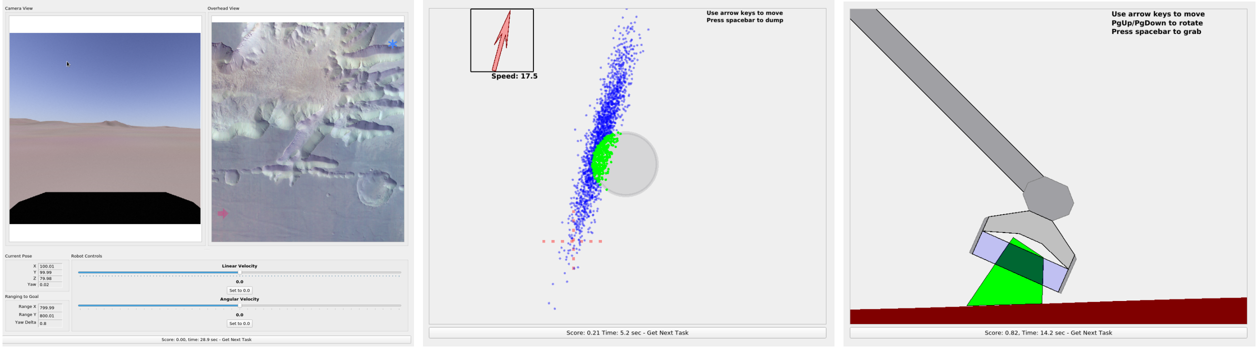


Fig. 3: *Relations between experimental controls, indicators of opportunity, and measures.*



X 100.01
Y 99.99
Z 79.98
Yaw 0.02

Range X 799.99
Range Y 800.01
Yaw Delta 0.8

Linear Velocity 0.0
Angular Velocity 0.0

Score: 0.00, time: 28.9 sec - Get Next Task

(a)

Use arrow keys to move
Press spacebar to dump

Score: 0.21 Time: 5.2 sec - Get Next Task

(b)

Use arrow keys to move
PgUp/PgDown to rotate
Press spacebar to grab

Score: 0.82, Time: 14.2 sec - Get Next Task

(c)

Fig. 4: *Experimental UI showing status information to the user in navigation (a), sample handling (b), and grasping (c) tasks.*

Using the *lme4* package in R [28], [29], we used a fully-dependent model of *productivity* with workload, attention, GSR, and HR terms, as well as an intercept term for actor and a random slope term for task type grouped by difficulty. Using Type II Wald chi-square tests to determine what factors are significant, we identified that attention alone ($\chi^2(6) = 15.8703$, p=0.014), and workload and attention together ($\chi^2(4) = 14.7087$, p=0.005) are significant predictors of productivity, giving us confidence that our experiment manipulated performance and answering Question 1a in the affirmative. Productivity is also significantly predicted by heartrate alone ($\chi^2(1) = 5.3067$, p=0.021), allowing us to offer heartrate as one candidate to help answer Question 1b. We also determined that individually modelling actors is critical to predicting productivity; comparing models with and without the intercept term for actor identity, we found a significant improvement, effect ($\chi^2(1) = 110.24$, p < 0.001), supporting a conclusion that individual differences are important to help answer Question 1c.

From our offline analysis of our experimental dataset, we can confidently conclude that significant relationships exist between our independent and certain of our dependent variables. Based on our analysis, we should be able to confidently predict performance given attention alone, workload and attention, or heartrate. However, we have only justified this result in a batch setting; our goal is to establish and exploit these relationships in an online setting. We have also not considered the effects of actor learning that may be present in our data.

IV. Assessment of a Reinforcement-Learning Adaptive Policy Using Human-Sourced Data

Using our pool of actor data from the experiment in Section III-B, we now assess our adaptive policy against more realistic data. Following Section III-A, we compare three assistant allocation policies – *random*, *linear multi-arm bandit*, and *informed static* – by their crossover point and final cumulative regret. The *random* policy assigns subtasks to actors with uniform probability over the set of all actors. No state information is used to inform the selection. The *linear multi-arm bandit* policy [21] estimates the potential reward at each time t for each actor assignment a as a linear combination of the current world ($\mathbf{z}$) and actor ($\mathbf{x}_a$) states with parameters β^* and θ^*_a, as $\mathbf{E}[r_{t,a} \mid \mathbf{x}_{t,a}, \mathbf{z}_t] = \mathbf{z}_t^T \beta^* + \mathbf{x}_{t,a}^T \theta^*_a$. Our experiment maps subtask difficulty to world state, and attention, workload, GSR, and heart rate to actor state, using a separate bandit for each subtask type. The estimates of β and θ_a are then updated with that actor's observed reward. This policy was chosen for its simplicity and expressive power and was validated in our prior work [1]. In the *informed static* policy is determined from a post-hoc analysis of the ground truth behaviour data. The informed static policy is a table which accumulates counts for the highest-earning task/actor pairing for each assignment. After processing all assignment events, the actor with the maximum count for each task is selected as the designated assistant. While this policy is not physically realizable, it acts as a baseline high-performance static policy, analogous to the static analysis proposed in [10]. This policy differs from the batch analysis in Section III-B, as only actor performance outcomes are considered.

Hypothesis 2: *An adaptive reinforcement learning policy that considers contextual information can outperform policies that do not consider such information, as measured by cumulative regret.* If accepted, we can conclude that our adaptive policy is useful in real-world situations in addition to the purely simulated environments of our previous work.

We conducted one hundred trials of each policy using subteams ranging from three actors in size to six actors in size, enumerated according to the minimally overlapping algorithm discussed in Section III-A. This is a 3x1 between-subjects design, where the 'subject' is the policy type.

Actor Count	**F Statistic**	p	η^2
3	F(2, 297) = 8.59	$p < 0.001$	0.0547
4	F(2, 297) = 7.12	$p = 0.001$	0.0458
5	F(2, 297) = 6.56	$p = 0.002$	0.0423
6	F(2, 297) = 7.01	$p = 0.001$	0.0451

TABLE I: *Hypothesis 3 ANOVA results, 150 selection events.*

A. Results

Fig. 5 shows the mean cumulative regret at each selection event for pools of three (a) and six (b) actors, zoomed in to highlight the crossover point signalling the transition from exploration to exploitation within the bandit policy. In each case, we observe that by the final selection event the bandit policy has, on average, achieved a lower cumulative regret than either a random or an informed static policy for all teams sizes, with ANOVA results reported in Table I. In each case, post-hoc tests using Tukey's HSD test confirmed that the bandit policy achieved a lower cumulative regret than either the informed static or random policies. The irregular distribution of the crossover points emphasize the importance of considering a set of trials when evaluating performance, rather than relying on singular examples. The histogram of crossover points for each condition is shown in Fig. 9 in the Supplementary Materials, as well as the full cumulative regret plots for all four conditions (Fig. 10 through Fig. 13).

In addition to observing the aggregate behavior over many trials, we examine several individual runs to show that the bandit is in fact learning. One such run is shown in Fig. 6, where a single trial's cumulative regret plot is shown. In this instance, the bandit makes a better choice than either of the two alternative policies. However, we caution that examining individual runs is not necessarily indicative of algorithmic performance, since discriminating between actual learning and dumb luck requires repeated trials to ensure that the algorithm makes superior choices over time.

B. Discussion

Our experiment in Section III-B manipulated subject performance, with statistical significance. While other assistant selection mechanisms use direct measurements of workload, our algorithm can unobtrusively infer workload from online observations of individuals. The statistical analyses show that in real subjects our individualized model is an improvement on the ideal static policy. Batch techniques such as the ideal static policy fail to consider performance as a function of contextual information, exhibiting similar performance to the random policy; such aggregate policies could be useful to initialize a learning system and reduce the number of needed training events.

As an example of the practical effects of our reward function definition, in Fig. 6 at the informed policy's choice at event 54, the informed policy incurred a regret of approximately 0.5 by choosing an actor requiring nearly twice as much time as the best choice of actor. This figure also highlights the importance of folding mission priorities into reward function design; the learning system will learn to maximize reward, potentially exploiting a poor design and yielding suboptimal outcomes.

In Fig. 5, the improvement from the multi-arm bandit policy is not as pronounced as in our prior simulated work. In fact, regret curves across all policies are similar, indicating that certain events had a wide variation in performance across subjects. One possible explanation for this is a lack of discriminatory power, since we used a feature vector in which all possible data were provided to the multi-arm bandit policy. Yet, our analyses show that not all data were useful in predicting productivity. In future work, we will investigate which contextual factors should be provided to selection algorithms, potentially including mission-relevant data such as location, bandwidth, available dexterity, and concurrent tasking. While our study population was not drawn from the highly educated and regimented population of astronauts, our use of a more general population shows that our method can be applied to a broad range of autonomy applications.

V. Conclusions

In this work, we presented an experimental design to gather a dataset upon which policies that solve the assistant selection problem can be tested without requiring large numbers of test subjects. In particular, we are interested in policies that exploit noninvasive indicators of opportunity to inform the selection process. We also presented the results of running this experiment and an initial analysis, finding that the use of such indicators yields an improvement in the quality of decisions, as measured by cumulative regret. We can therefore assess that our adaptive reinforcement-learning based approach is useful with real humans and warrants further study.

While our methods can still produce improvements over policies that do not use contextual information to make decisions under uncertainty, our data emphasize the need for further investigation into the most informative observations that yield the most predictive power with respect to estimating performance. We are also interested in the role of the learning process, exploring how an adaptive approach can be modified to model the changes in a human actor's performance as a result of becoming more familiar with the limits of a robotic system. Our work is applicable to any imperfect autonomous system that requires human intervention; while this work has focused on space robotics, we may readily apply our technique to other fields such as robotic agriculture or self-driving transportation.

References

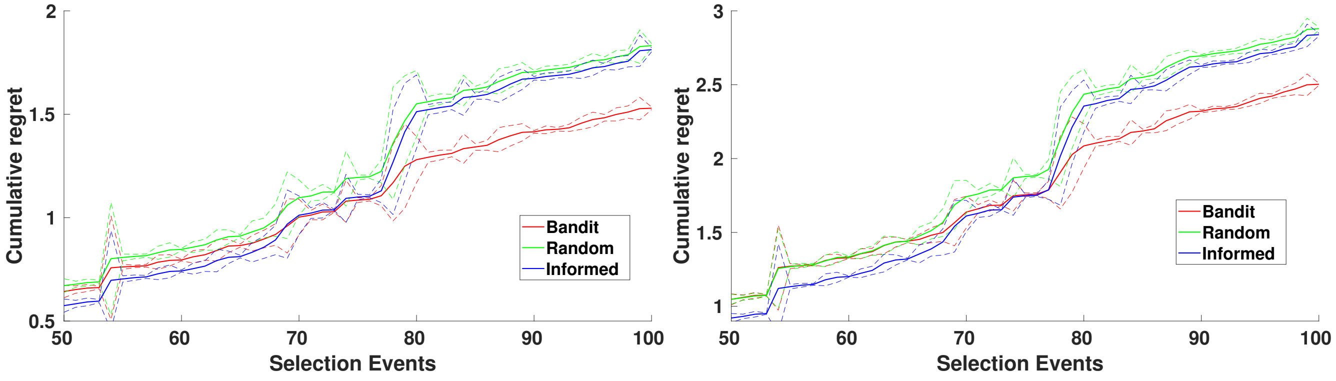
[1] S. McGuire, P. M. Furlong, C. Heckman, S. Julier, D. Szafir, and N. Ahmed, "Failure is not an option: Policy learning for adaptive recovery in space operations," *IEEE Robotics and Automation Letters*, vol. 3, no. 3, pp. 1639–1646, 2018.

[2] T. Fong, M. Allan, X. Bouyssounouse, M. G. Bualat, M. Deans, L. Edwards, L. Flückiger, L. Keely, S. Y. Lee, D. Lees, *et al.*, "Robotic site survey at Haughton Crater," 2008.

[3] B. G. Drake and e. Watts, Kevin D, "Human Exploration of Mars Design Reference Architecture 5.0, Addendum #2," *NASA/SP-2009-566-ADD2*, 2014.

[4] A. Elfes, C. R. Weisbin, H. Hua, J. H. Smith, J. Mrozinski, and K. Shelton, "The HURON task allocation and scheduling system: Planning human and robot activities for lunar missions," in *Automation Congress, 2008. WAC 2008. World.* IEEE, 2008, pp. 1–8.

[5] B. Sankaran, B. Pitzer, and S. Osentoski, "Failure recovery with shared autonomy," in *Intelligent Robots and Systems (IROS), 2012 IEEE/RSJ International Conference on.* IEEE, 2012, pp. 349–355.



(a) 3 actor condition

(b) 6 actor condition

Fig. 5: *Plots of cumulative regret for two actor conditions; 2σ bounds in dashed lines. Results have been zoomed to highlight crossover between bandit policy and informed static policy. Cumulative regret is expressed in units of productivity, the ratio between score earned and time required for each event.*

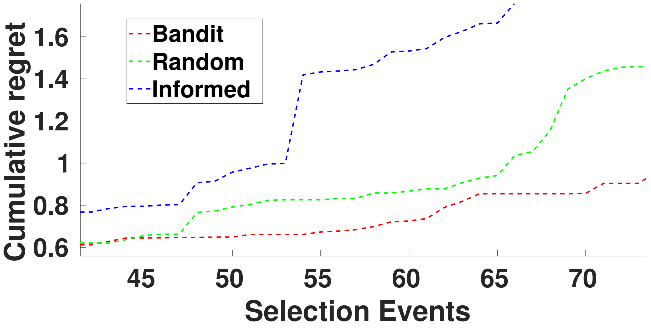


Fig. 6: *Individual run of a policy selection among three actors, showing good choices by the bandit policy at selection events 48 and 66*

[6] R. A. Knepper, S. Tellex, A. Li, N. Roy, and D. Rus, "Recovering from failure by asking for help," *Autonomous Robots*, vol. 39, no. 3, pp. 347–362, 2015.

[7] V. Verma, G. Gordon, R. Simmons, and S. Thrun, "Real-time fault diagnosis," *IEEE Robotics & Automation Magazine*, vol. 11, no. 2, pp. 56–66, 2004.

[8] T. Fong, C. Thorpe, and C. Baur, "Robot, asker of questions," *Robotics and Autonomous systems*, vol. 42, no. 3, pp. 235–243, 2003.

[9] D. Schreckenghost, T. Milam, and T. Fong, "Measuring performance in real time during remote human-robot operations with adjustable autonomy," *IEEE Intelligent Systems*, vol. 25, no. 5, pp. 36–45, 2010.

[10] M. Johnson, J. M. Bradshaw, P. J. Feltovich, C. M. Jonker, M. B. Van Riemsdijk, and M. Sierhuis, "Coactive design: Designing support for interdependence in joint activity," *Journal of Human-Robot Interaction, 3 (1), 2014*, 2014.

[11] M. A. Goodrich and D. R. Olsen, "Seven principles of efficient human robot interaction," in *Systems, Man and Cybernetics, 2003. IEEE International Conference on*, vol. 4. IEEE, 2003, pp. 3942–3948.

[12] H. W. Kuhn, "The Hungarian method for the assignment problem," *Naval research logistics quarterly*, vol. 2, no. 1-2, pp. 83–97, 1955.

[13] B. P. Gerkey and M. J. Mataric, "Sold!: Auction methods for multirobot coordination," *IEEE transactions on robotics and automation*, vol. 18, no. 5, pp. 758–768, 2002.

[14] D. P. Bertsekas, "The auction algorithm for assignment and other network flow problems: A tutorial," *Interfaces*, vol. 20, no. 4, pp. 133–149, 1990.

[15] D. Zhu, H. Huang, and S. X. Yang, "Dynamic task assignment and path planning of multi-AUV system based on an improved self-organizing map and velocity synthesis method in three-dimensional underwater workspace," *IEEE Transactions on Cybernetics*, vol. 43, no. 2, pp. 504–514, 2013.

[16] L. F. Bertuccelli and J. P. How, "Active exploration in robust unmanned vehicle task assignment," *Journal of Aerospace Computing, Information, and Communication*, vol. 8, no. 8, pp. 250–268, 2011.

[17] A.-B. Karami, L. Jeanpierre, and A.-I. Mouaddib, "Partially observable Markov decision process for managing robot collaboration with human," in *Tools with Artificial Intelligence, 2009. ICTAI'09. 21st International Conference on*. IEEE, 2009, pp. 518–521.

[18] J. Barry, L. P. Kaelbling, and T. Lozano-Pérez, "Hierarchical solution of large Markov decision processes," 2010.

[19] L. E. Parker, "Task-oriented multi-robot learning in behavior-based systems," in *Intelligent Robots and Systems' 96, IROS 96, Proceedings of the 1996 IEEE/RSJ International Conference on*, vol. 3. IEEE, 1996, pp. 1478–1487.

[20] R. Francis, T. Estlin, G. Doran, S. Johnstone, D. Gaines, V. Verma, M. Burl, J. Frydenvang, S. Montaño, R. Wiens, *et al.*, "Aegis autonomous targeting for chemcam on mars science laboratory: Deployment and results of initial science team use," *Science Robotics*, vol. 2, no. 7, p. eaan4582, 2017.

[21] L. Li, W. Chu, J. Langford, and R. E. Schapire, "A contextual-bandit approach to personalized news article recommendation," in *Proceedings of the 19th International Conference on World Wide Web*. ACM, 2010, pp. 661–670.

[22] O. Cappé, A. Garivier, O.-A. Maillard, R. Munos, G. Stoltz, *et al.*, "Kullback–Leibler upper confidence bounds for optimal sequential allocation," *The Annals of Statistics*, vol. 41, no. 3, pp. 1516–1541, 2013.

[23] A. P. Burger, W. Grundlingh, and J. H. van Vuuren, "Two combinatorial problems involving lottery schemes: Algorithmic determination of solution sets," *Utilitas Mathematica*, vol. 70, pp. 33–70, 2006.

[24] W. H. Kruskal and W. A. Wallis, "Use of ranks in one-criterion variance analysis," *Journal of the American Statistical Association*, vol. 47, no. 260, pp. 583–621, 1952.

[25] J. Heard, C. E. Harriott, and J. A. Adams, "A human workload assessment algorithm for collaborative human-machine teams," in *Robot and Human Interactive Communication (RO-MAN), 2017 26th IEEE International Symposium on*. IEEE, 2017, pp. 366–371.

[26] J. T. Cacioppo and L. G. Tassinary, "Inferring psychological significance from physiological signals." *American Psychologist*, vol. 45, no. 1, p. 16, 1990.

[27] Y. Santiago-Espada, R. R. Myer, K. A. Latorella, and J. R. Comstock Jr, "The multi-attribute task battery ii (matb-ii) software for human performance and workload research: A user's guide," 2011.

[28] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2018. [Online]. Available: https://www.R-project.org/

[29] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *Journal of Statistical Software*, vol. 67, no. 1, pp. 1–48, 2015.

Supplementary Materials

We have included several plots in these materials that while not critical to our work, provide a more complete picture of the data and analyses involved.

Histograms from Section III-A, Synthesizing from Small n

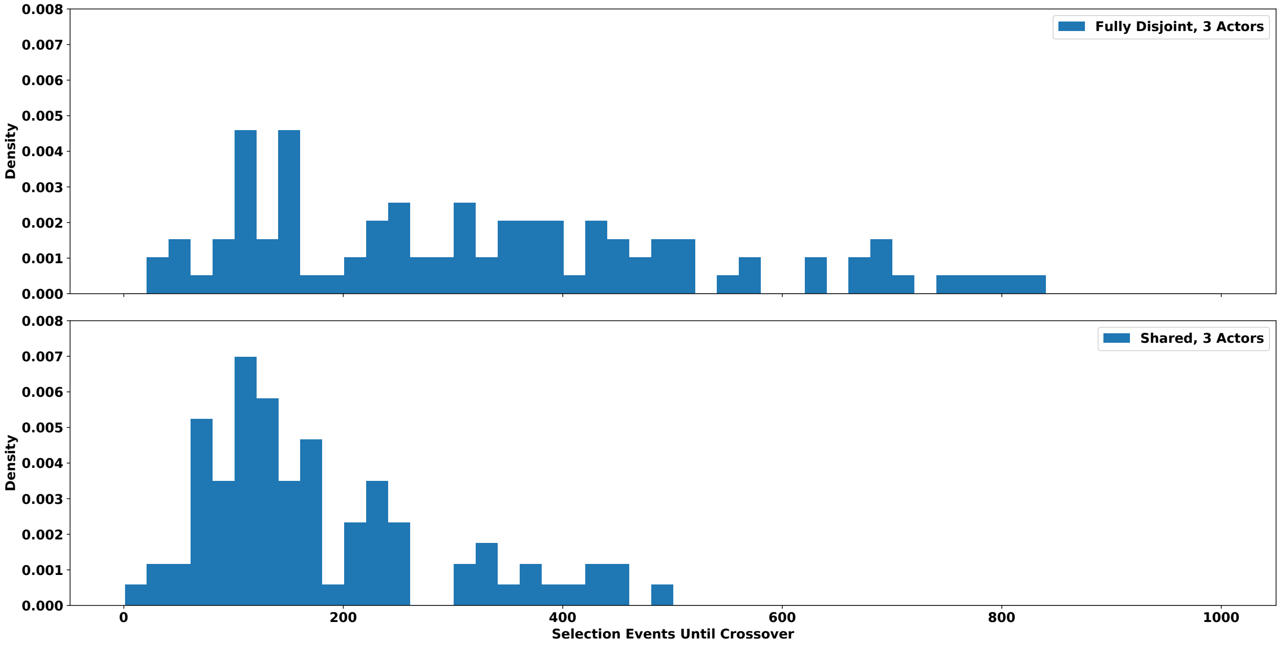


Fig. 7: Distribution of crossover points for 3 actors

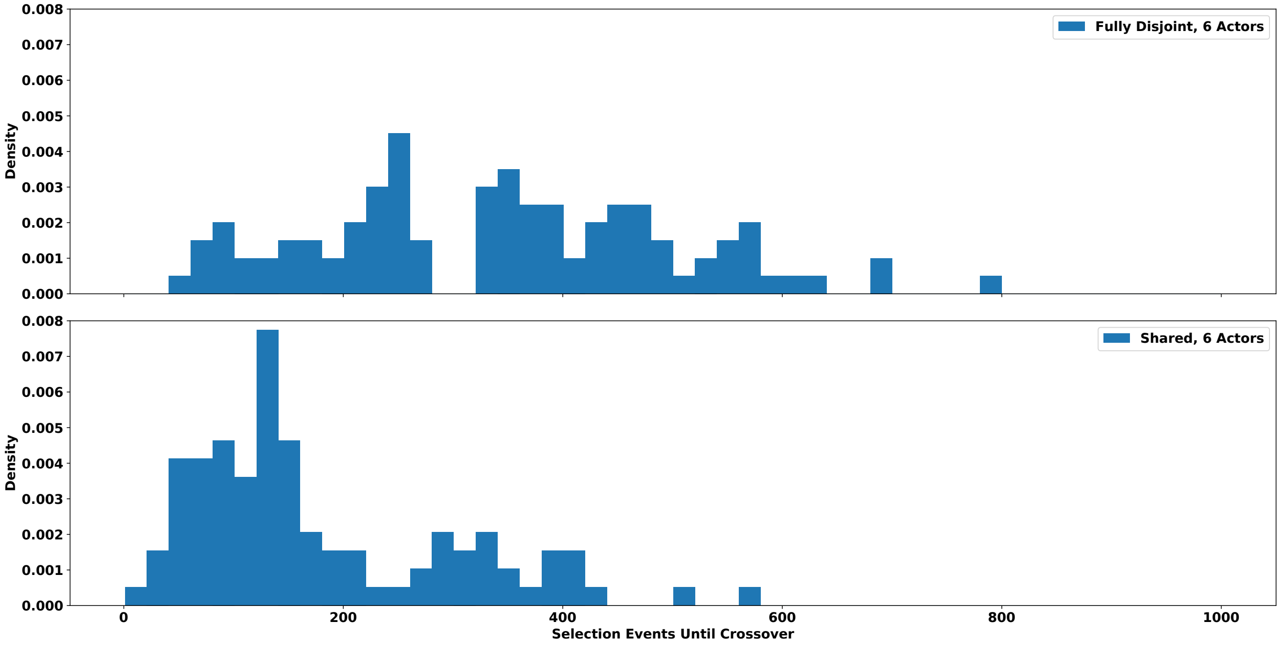


Fig. 8: Distribution of crossover points for 6 actors

Histograms from Section IV: Assessment of a reinforcement-learning adaptive policy using human-sourced data

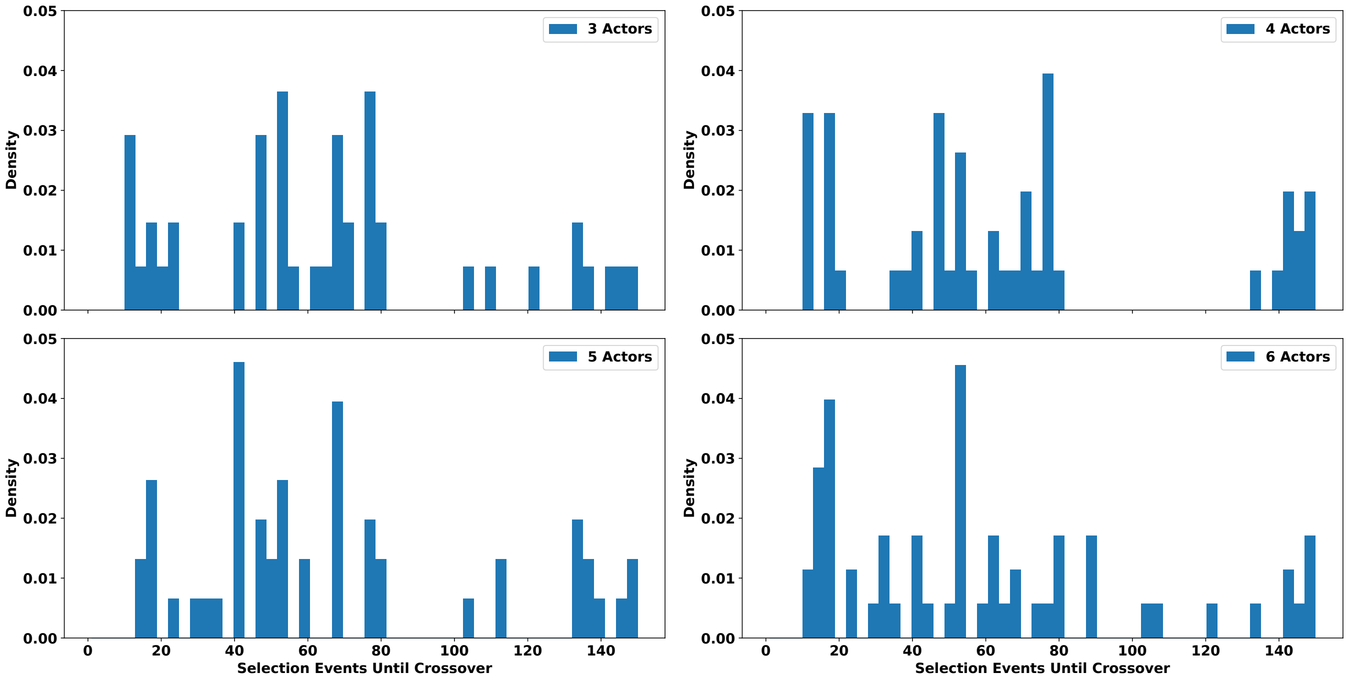


Fig. 9: Histograms of crossover points, showing the relative density of the number of selection events needed before a bandit policy outperforms an informed static policy

Cumulative Regret plots from Section IV: Assessment of a reinforcement-learning adaptive policy using human-sourced data

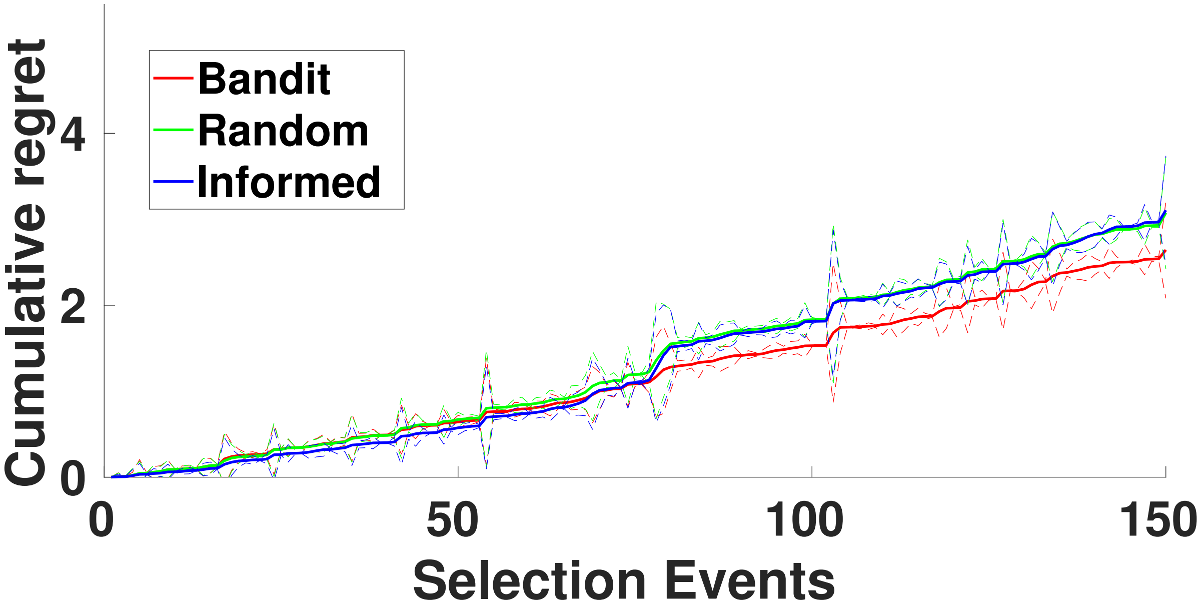


Fig. 10: Plots of cumulative regret for 3 actors; $\pm 5\sigma$ bounds are shown in dashed lines

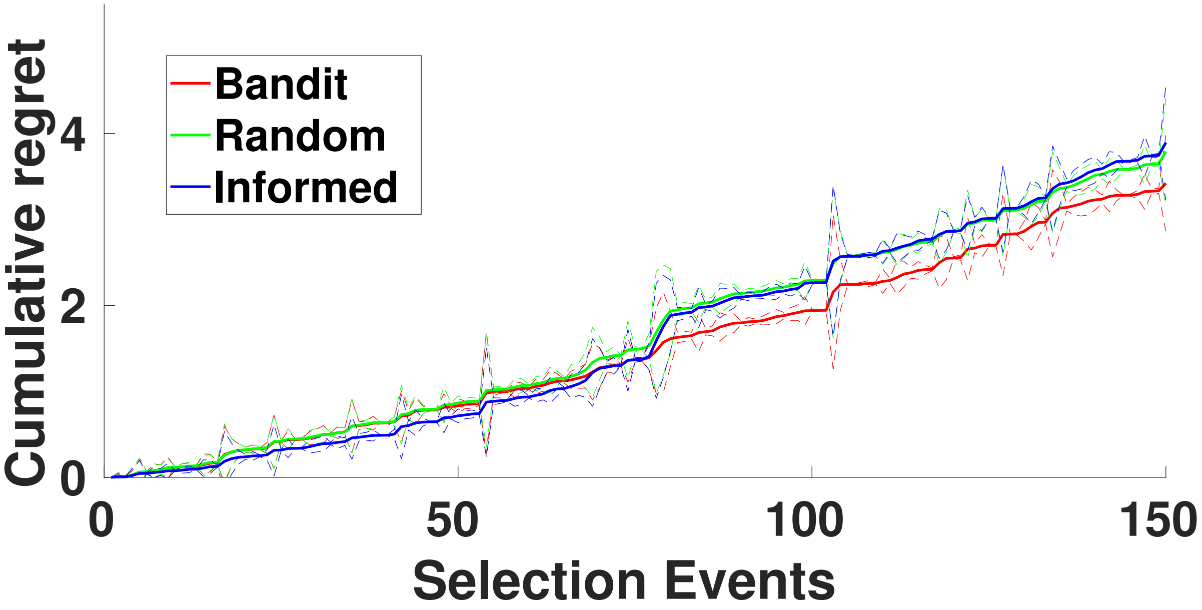


Fig. 11: Plots of cumulative regret for 4 actors; $\pm 5\sigma$ bounds are shown in dashed lines

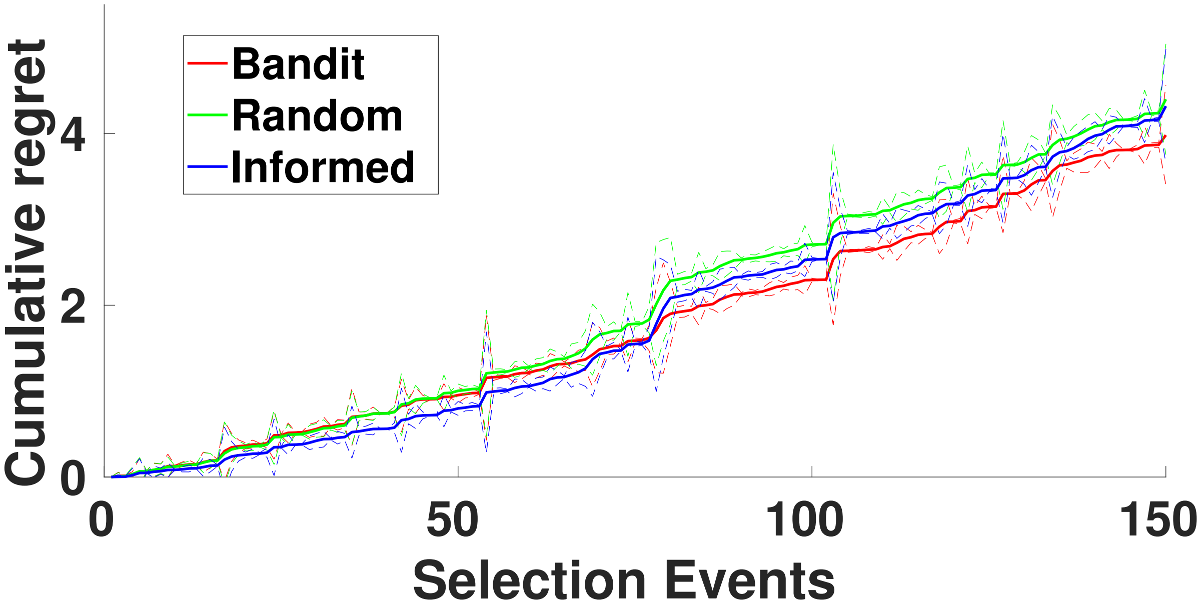


Fig. 12: Plots of cumulative regret for 5 actors; ±5σ bounds are shown in dashed lines

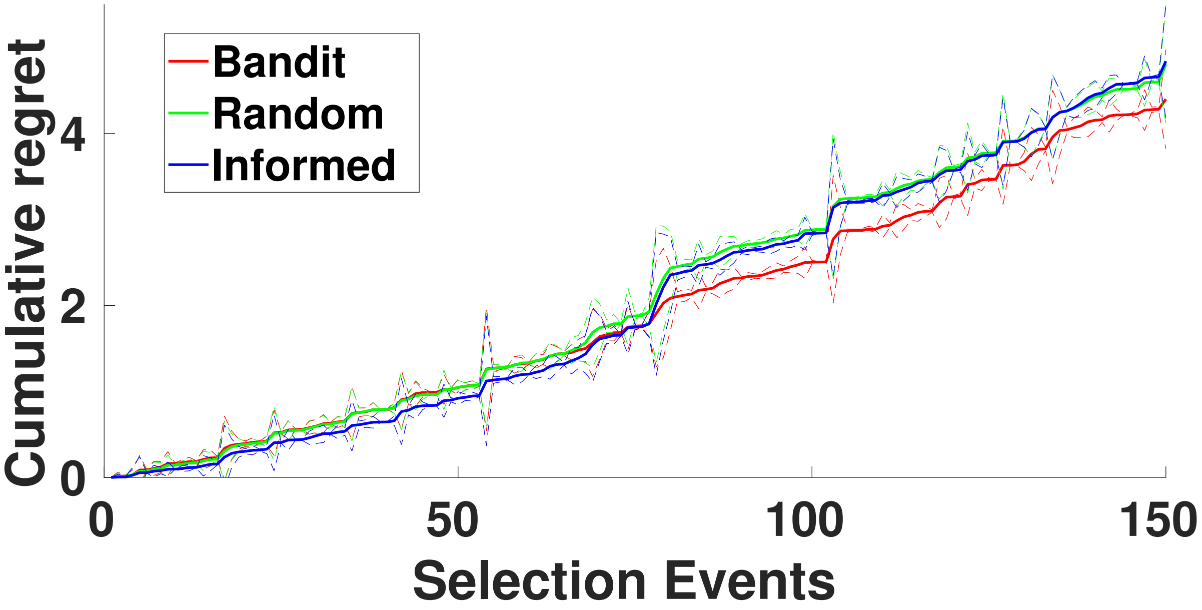


Fig. 13: Plots of cumulative regret for 6 actors; ±5σ bounds are shown in dashed lines

Analysis of Deviance Table (Type II Wald χ^2 tests)
Model: $Productivity \sim$ $Workload * Attention * GSR * HR +$
$(1 + TaskType/TaskDifficulty) + (1|Actor)$

Response: Productivity			
	χ^2	Df	Pr($> \chi^2$)
Workload	4.9634	6	0.548510
Attention	15.8703	6	0.014468
GSR	0.0079	1	0.929395
HR	5.3067	1	0.021244
TaskType	1789.4987	2	$<$ 2.2e-16
Workload:Attention	14.7087	4	0.005345
Workload:GSR	2.7055	2	0.258528
Attention:GSR	5.3802	2	0.067874
Workload:HR	2.3334	2	0.311388
Attention:HR	1.0852	2	0.581230
GSR:HR	0.2499	1	0.617142
TaskType:TaskDifficulty	1037.6611	6	$<$ 2.2e-16
Workload:Attention:GSR	2.1026	4	0.716899
Workload:Attention:HR	4.6067	4	0.330079
Workload:GSR:HR	2.0478	2	0.359183
Attention:GSR:HR	3.2907	2	0.192945
Workload:Attention:GSR:HR	3.7095	4	0.446752

TABLE II: Type II Wald Test Results. Rows indicating observable signals of interest are bolded, while rows indicating experimental controls are italicized. Heartrate (HR) alone is a significant predictors of performance, while neither galvanic skin response (GSR) nor GSR with HR are significant predictors of performance. The prediction of the italicized control variables indicates that our experimental controls were effective in manipulating performance.