Check for updates

A User Study Evaluation of Predictive Formal Modelling at Runtime in Human-Swarm Interaction

AYODEJI O. ABIOYE, University of Southampton, UK and The Open University, UK
WILLIAM HUNT, University of Southampton, UK
YUE GU, University of Glasgow, UK
EIKE SCHNEIDERS, University of Southampton, UK
MOHAMMAD NAISEH, Bournemouth University, UK
BLAIR ARCHIBALD, University of Glasgow, UK
MICHELE SEVEGNANI, University of Glasgow, UK
SARVAPALI D. RAMCHURN, University of Southampton, UK
JOEL E. FISCHER, University of Nottingham, UK
MOHAMMAD D. SOORATI, University of Southampton, UK

Formal Modelling is often used as part of the design and testing process of software development to ensure that components operate within suitable bounds even in unexpected circumstances. We conducted a user study evaluation of predictive formal modelling (PFM) at runtime in a human-swarm mission to determine the benefit of predictive formal modelling on performance and human-swarm interaction. 180 participants were recruited to perform the role of aerial swarm operators delivering parcels to target locations in a simulation environment. The PFM model was integrated into the simulation software to inform the operator of the estimated mission completion time given the current number of drones deployed. The operator could increase the number of parcels delivered in any time step by adding drones, which also increased costs, thus requiring the use of the minimum number of drones necessary to complete the task in the given time. We collected user feedback using standard survey questionnaires and measured performance using data obtained from the Human And Robot Interactive Swarm (HARIS) simulator. Our results show that PFM increased the performance of the human swarm team without significantly increasing the operators' workload or affecting the system's usability.

CCS Concepts: • **Human-centered computing → Empirical studies in HCI**; • **Computing methodologies → Model verification and validation**.

Additional Key Words and Phrases: Human-Robot Interaction (HRI), Human-Swarm Interaction (HSI), Predictive Formal Modelling (PFM), Workload, Usability

Authors' Contact Information: Ayodeji O. Abioye, a.o.abioye@soton.ac.uk, ayodeji.abioye@open.ac.uk, University of Southampton, Southampton, UK and The Open University, Milton Keynes, UK; William Hunt, w.hunt@soton.ac.uk, University of Southampton, Southampton, UK; Yue Gu, yue.gu@glasgow.ac.uk, University of Glasgow, Glasgow, UK; Eike Schneiders, eike.schneiders@soton.ac.uk, University of Southampton, Southampton, UK; Mohammad Naiseh, mnaiseh1@bournemouth.ac.uk, Bournemouth University, Bournemouth, UK; Blair Archibald, blair.archibald@glasgow.ac.uk, University of Glasgow, Glasgow, UK; Michele Sevegnani, michele.sevegnani@glasgow.ac.uk, University of Glasgow, Glasgow, UK; Sarvapali D. Ramchurn, sdr1@soton.ac.uk, University of Southampton, Southampton, UK; Joel E. Fischer, joel.fischer@nottingham.ac.uk, University of Nottingham, Nottingham, UK; Mohammad D. Soorati, m.soorati@soton.ac.uk, University of Southampton, Southampton, UK.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
© 2025 Copyright held by the owner/author(s).
ACM 2573-9522/2025/4-ART
https://doi.org/10.1145/3727989

Just Accepted

1 INTRODUCTION

Aerial swarms have the potential to change how we conduct disaster management. In search and rescue operations, they can be deployed to locate and identify casualties faster than a single drone or manual search by human teams. Aerial swarms can be used to deliver life-saving supplies to remote and difficult-to-reach areas simultaneously in the aftermath of a natural disaster. Therefore, aerial swarms have the potential to be an essential part of the human response team. Such operations will consist of human-swarm teams working together to achieve mission success. However, a major challenge with deploying aerial swarms is the issue of resource allocation. In practical operations, the human swarm team will have only a limited number of unmanned aerial vehicles (UAVs) available to deploy in their mission. Therefore, knowing exactly how many UAVs are needed for the mission to succeed at any point in time is crucial. This ensures the efficient distribution of the limited resources to where they are needed in advance or the early re-tasking of some resources during the mission in real-time as events unfold or as the mission conditions change.

Studies in human-swarm interaction (HSI) have identified essential prerequisites for the successful operation of aerial swarms [16]. For systems relying on human supervision and intervention, a critical requirement for the smooth operation of the swarm is the efficient timing and selection of relevant data provided to the operator. [23] proposed a predictive formal modelling (PFM) technique to estimate the mission success at runtime. PFM can be used to inform the swarm operator about the crucial information required to make appropriate decisions in order for the mission to be successful. In this paper, we integrate PFM into the Human And Robot Interactive Swarm (HARIS) simulator [25] to provide human swarm operators with real-time mission and swarm status updates, along with predictions of mission success. The goal is to assess the usability, impact on workload, and impact on performance as a result of the added information provided by the PFM feature via a user study. We previously showed that PFM increased performance without affecting workload or the system's usability through a user study with 60 participants [2]. We extended our previous research by recruiting another 120 participants and introduced an alternative prediction model, the bounded random prediction model (BRPM) to further test and validate our findings and to understand the impact of PFM on human-swarm interaction. We conducted more experiments focusing on two new comparisons. Firstly, BRPM was compared with the No-PFM condition using 60 participants. Secondly, we compared the BRPM condition against the PFM condition with the remaining 60 participants.

The next section of this paper discusses relevant related works in human-swarm interaction, swarm verification, probabilistic model checking, and multi-agent pickup and delivery. The system model section presents the two prediction models PFM and BRPM. The methodology section describes our user study and consists of subsections focused on the study scenario, task, and procedure. The next section after this is the result and analysis section. The results are presented in three parts: the result of Experiment 1, Experiment 2a, and Experiment 2b. This is followed by a discussion of the major findings and then the conclusion.

2 RELATED WORK

2.1 Human-Swarm Interaction

The combined factors of the limited ability of autonomous robots to perform complex organisational tasks within a swarm, and the moral and multifaceted decisions that must be made during a search and rescue operation make it appealing to pursue a human-in-the-loop system; one that has a human (or humans) continuously involved with the swarm. This can take multiple forms such as direct supervisory control via tasking agents [29]; control of leader agent that others follow [53]; abstract influence via beacons placed in the environment [29]. Including a human in the decision-making process of an autonomous collective often improves performance and speed [17]. In most cases, the intention is to reduce the perceived complexity of the swarm via interface data or external information such that the human can perceive the swarm as a single entity, thus their cognitive

Just Accepted

workload (which typically scales with the number of agents) is controlled. This would allow a single human to control a large number of robots in a manner that would otherwise be impractical, offering the advantage of human decision-making and moral oversight at every level of the operation [30].

In one related study, Abioye et al. [3] investigated how swarm operators' performance is affected when the user interface presents low-quality data compared to high-quality data due to communication issues. The study showed no significant difference in the two conditions in terms of the operator's performance, however, the operator's trust was significantly higher in the high-quality data condition. Schneiders et al. [48] indicated the demand for studying non-dyadic human-in-the-loop system configurations, such as that presented in this work. Hunt et al. [25] proposed a method of dynamic re-tasking and triage based on operator feedback as well as the HARIS simulator, a browser-based platform that was specifically designed for human-in-the-loop multi-agent and swarm robotics experimentation. HARIS is a successor of HutSim [45] which was designed with a specific focus on usability by consulting with industry experts to model not only their typical command structure but also make it operable as a digital twin with multiple operators and real-life or simulated UAVs [49]. Building on its predecessor, HARIS was further tailored to its use case derived from interviews with drone pilots [44] and swarm experts [42] to make the platform as usable and realistic as possible while maximising the ease of use for multiple human operators [49], making this simulator a useful tool for the investigations on human-swarm simulations. We extend the existing work by investigating the effect of providing the operator with the estimated completion time and cost based on a predictive formal model.

In addition, several studies discussed the role of calibrated trust for safe and effective human-swarm partnerships. For instance, in [28], the researchers pointed out a few incidents where undertrust of the operators towards the autonomous systems and how they led to catastrophes. On the contrary, over-trust can lead users to favour robots over human intuition or recommendation which can also be dangerous [9, 46]. These results call for designing human-swarm interfaces with a calibrated trust goal in mind [41]. Another important factor in human-swarm interaction is measuring operator workload. The NASA TLX [24] is a common tool used to assess workload in human-swarm interaction [12] and in human-robot interaction research [55]. The NASA TLX method of measuring workload agrees with other objective methods such as EEG [40]. Therefore, the NASA TLX was used in this study to measure workload.

Just Accepted

2.2 Swarm Verification

Formal methods, e.g. model checking, have been applied in previous studies to verify swarm systems, such as the satisfaction of emergent behaviours [32], the analysis of security requirements in unbounded swarms [10] and the reasoning about fault tolerance in probabilistic systems [37]. However, none of these approaches can give guarantees after deployment. A close approach to ours is runtime monitoring [5], where pre-constructed monitors are used to analyse the system execution traces that are generated at runtime against formal specifications [20]. These monitors can evolve with system dynamics, such as the size and topology, but cannot reason about swarm-level specifications, like the human-swarm interactions, where finite observations are not sufficient. Instead, Gu et al. [23] propose a framework to integrate runtime modelling [8], that has been deployed in system reasoning for unforeseen situations during execution [7], with formal methods and focus on formal runtime modelling. Quantitative formal models can provide predictions, and this has been used at design time, e.g. for predicting failures and service availability of components [14]. In this work, we improve the accuracy and efficiency of an existing model [23] and implement PFM at runtime, to predict the feasibility of human-swarm missions (i.e. estimated completion time) and evaluate its effect on calibrating the trust of human operators.

2.3 Probabilistic Model Checking

Similar to [23], our swarm scenarios are modeled as (labelled) continuous-time Markov chains (CTMCs) which are analysed through probabilistic model checking. A CTMC is a triple: $C = (S, R, L)$, where S is a finite set of states with a designated initial state; $R : S \times S \rightarrow \mathbb{R}_{\geq 0}$ is a rate matrix and consists of a set of non-negative real numbers; and $L : S \rightarrow 2^{AP}$ is a labelling function and assigns to each state in S a set of $L(s)$ of (A)tomic (P)roposition (AP) that are valid in the state, e.g. $s \mapsto \{\text{failure}\}$. The transition rate matrix R assigns rates to each pair of states in a CTMC and represents the likelihood of transition between these two states in real time. A transition can only occur between s and s' if $R(s, s') > 0$ and, by default, the probability of this transition being triggered within t time units satisfies an exponential distribution and equals to $1 - e^{-R(s,s') \cdot t}$.

We build our CTMC models using the PRISM modelling language [33] that provides a model checker to quantify *all* possible system behaviours, e.g. querying the probability a system will succeed. Such properties are specified in Continuous Stochastic Logic (CSL) [4]. CSL is a temporal logic with probabilistic operators including the *eventually* $\mathbf{F}\,\varphi$ temporal operator that specifies, for all paths, we eventually reach a state where φ is true. In PRISM, properties can be quantified through an operator $\mathcal{P}_{=?}\,[\,\Psi\,]$. It determines the *likelihood* that a path exists where Ψ is true. As suggested in [23], for large swarm models, *statistical model checking* (SMC) [54] can effectively sample the model space through repeated simulation resulting in a timely performance without a significant cost to accuracy. SMC is supported by the built-in discrete-event simulator in PRISM.

2.4 Multi-Agent Pickup and Delivery

An important emerging application of multi-agent systems research is that of package delivery — using autonomous robots such as UAVs to transport packages from a local warehouse to their final destination (the "last mile") [6]. Such a scenario can be formulated either *offline*, where the task and agent sets are known from the outset [36], or *online* where the new agents, or more typically new tasks, are added at runtime and the swarm must accommodate and re-plan on-the-fly [15]. Developments in this area include careful design of the environment such as the interior of the warehouse in which these agents operate [47] as well as dynamic transfer of packages between agents [38]. Other works have included hybrid human-robot teams in warehouses, the "cobot" model, where robots must carefully work around humans without causing accidents [22]. Autonomy of agent behaviour is common in almost all works in package delivery, making it an appropriate usecase for a human operator monitoring the swarm and having a more abstract influence on the mission than they might in other tasks.

In this work we model a package delivery problem where the operator monitors a small swarm of agents which must perform 40 delivery tasks. The agents autonomously allocate themselves, collect packages from the hub where they start, and deliver them to the task locations. Each agent can only carry one package at a time, and agents each have a battery which decays and causes them to return to the hub to recharge if critically depleted. The operator can observe the mission and can alter the size of the swarm by adding or removing drones using two buttons. When the user clicks to add a drone, a new drone appears at the hub and allocates itself. When the user clicks to remove a drone, the next drone to reach the hub is removed. The mission is successfully completed when every package has been delivered.

3 SYSTEM MODEL

3.1 Predictive Formal Model

We adapt the mission feasibility model of [23] to determine the probability of success for a set of delivery tasks and a group of UAVs. The model uses Continuous-Time Markov Chains (CTMC), and is built using the PRISM modelling language. Key elements of the model include:

- The 2D environment is discretised into 8 spatial regions, with UAVs always in a specific region.
- UAVs move through regions to their task location, and then follow a similar route back to the hub.

Just Accepted

- Each UAV's battery level is discretised into 5 values: full, high, mid, low and critical. UAVs with critical batteries need to be recharged at the hub before re-joining the group.
- All parameters, including UAV movement rate, battery draining rates outside the hub and the recharging rate in the hub, are aligned with the HARIS simulator parameters.

Figure 1 illustrates the different elements of the final CTMC[1]. When a UAV is assigned a task, it moves upwards through regions towards the task location. On the way back, it rotates and moves downwards through regions (see Figure 1a). While UAVs are outside the hub, their battery drains from *full* → *high* → *mid* → *low* → *critical*. A UAV with *critical* battery needs to return to the hub to recharge (see Figure 1b). The runtime implementation allows the initial states, including the initial locations and battery levels, to be the current states of UAVs. In other words, the Markov process can start from any state and transitions will follow the same manner.

We have tailored the original model to fit our user study. The background failure in each region is removed to relieve participants' stress from the unexpected loss of UAVs. We conducted pre-studies and found the distribution of the UAV dynamics in real-time is not exponential (as a CTMC assumes). For a more realistic model, we use the Erlang distribution and implement the Erlang-k law [19] to represent a smooth transition delay in CTMCs. It is obtained by reforming the state transition to a series of k-step transitions with the exponential rate $\frac{k}{t}$, where t is the transition time. As suggested in [13], we use $k = 4$, as illustrated in Figure 1c, for a good compromise between an accurate approximation and the computational overhead introduced by this technique.

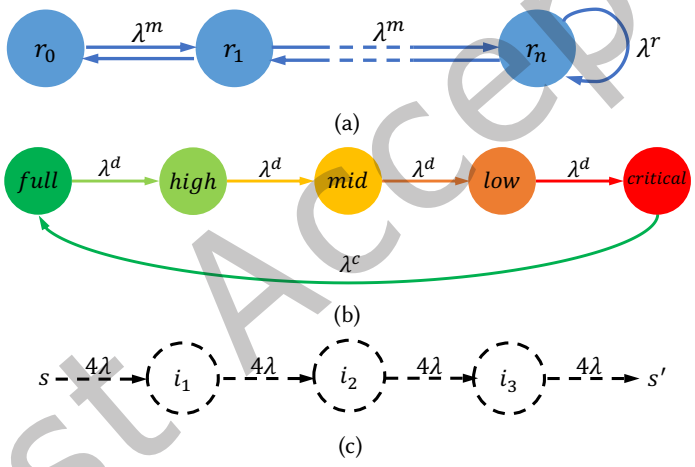


Fig. 1. Partial CTMC: (a) the UAV movement is modelled as transitions between regions r_n, where $n \in [0, 8]$ and r_0 is the hub, with corresponding movement rates λ^m and the rotation rate λ^r; (b) the battery drainage is synchronised with draining rates λ^d outside the hub and the charging rate λ^c in the hub; (c) for each transition $s \xrightarrow{\lambda} s'$ in (a) and (b), intermediate states $i_{1 \to k-1}$ are introduced to implement the Erlang-k law with transition rates $k \cdot \lambda$. In this case, we use $k = 4$.

We use the following time-bounded property to predict mission feasibility:

$$\mathcal{P}_{=?} \left[\mathbf{F}^{\leq T} \, all\, delivered \right] \tag{1}$$

It calculates ($_{=?}$) the probability ($\mathcal{P}$) that all delivery tasks are accomplished eventually ($\mathbf{F}$) within time T. We apply SMC to approximate the probability through repeated simulation of the mission feasibility model. A previous study has shown that SMC can achieve a timely prediction without a significant cost to accuracy [23].

[1]The final CTMC combines and synchronises these elements to create a single state space with appropriate transition rates.

Just Accepted

We integrate our (meta) model with HARIS to perform modelling at runtime as shown in Figure 2. This follows a similar process to [23], but with the Sim2PRISM middleware directly embedded in HARIS. As the simulation runs, HARIS constructs the model instantiations directly and, in parallel, calls PRISM to run the analysis. Instead of showing the probability of success directly, which might be difficult for participants to interpret, we consider the feasibility over different time intervals (i.e. T in Eq. 1) and give an estimated completion time as the time when the probability of success reaches 0.99, as shown below.

$$t_{pred} = t_{current} + T_{\mathcal{P}=0.99} \tag{2}$$

where t_{pred} is the completion time predicted by the PFM model; $t_{current}$ is the current time; $T_{\mathcal{P}=0.99}$ is the exact time delay from the current time when the probability of success reaches 0.99. Additionally, we implement *what-if* scenarios and follow the same calculation to give the participants extra information on the effect of adding/removing a UAV before making a decision.

3.2 Bounded Random Predictive Model

The introduction of an assistive interface component may result in a placebo effect, where the user's belief in the efficacy of the assistive tool increases their confidence, trust, or possibly their performance [31]. The prevalence of this effect in interactive interfaces is well established [52]. In our previous work, we compared only PFM and no PFM [2], and so we developed a Bounded Random Predictive Model (BRPM) to compare the impact of inaccurate and misleading predictions on the objective performance and subjective experience of the human user, when compared with PFM. This method does not use formal modelling and instead randomly predicts, giving similar-looking results that are erroneous. This allows us to test the impacts of a "good" model compared to a "dummy" one, and delineate the effects of PFM versus the effects of any indicator's presence. The equation describing the BRPM model is given below:

$$t_{pred} = t_{current} + T_{rand}, \quad T_{rand} \in [200, 400] \tag{3}$$

The predicted completion time t_{pred} of the BRPM is a random time T_{rand} uniformly sampled from interval $[200, 400]$ seconds and added to the current time $t_{current}$. This means that each time the model is used to predict the completion time, a different random t_{pred} is chosen, and added to the $t_{current}$ when it is called. This means that adding and removing drones, or new updates being triggered by time has a completely random effect on the prediction, so the user does not typically see the estimated completion time reduce when a drone is added, or vice versa. BPRM always predicts some time (minimum 200 seconds) into the future, and never sooner. Even if there is only one task remaining, the model will erroneously predict several minutes until completion.

This provides initially believable results to the user that are commensurate with the actual model. As the experiment goes on, BRPM estimates erratically into the future. The expectation is that users will believe this prediction, causing them to potentially add unnecessary drones or remove required ones. After a short while it will become increasingly clear to an attentive user that their actions are having no effect on the predictions, and that the predicted finish time is creeping into the future. At this point, if the users are over-trusting the model, they will be likely to fail, as they believe the incorrect prediction. However, if the users are cautiously using the model but paying attention to its updates and the situation on the map, then they will likely lose trust in the prediction and ignore it.

4 METHODOLOGY

We conducted three within-subject user studies, each with 60 participants as shown in Table 1 and Table 2. The participants in each experiment were divided into two counterbalanced groups, A and B, to eliminate the learning effect. We recruited a total of 180 participants (113 female, 65 male, 2 non-binary, average age: 35.9, all

Just Accepted

participants were above 18 years old). All participants were recruited through Prolific[2][3]. 63% of participants had at least a bachelor's degree, 71% reported average or above computer expertise, and 31% were familiar with UAV or swarm robotics (although only 3% reported first-hand experience). As the tutorial, the scenarios, as well as the questionnaires were presented in English, all participants were recruited from the US and the UK. Participants were rewarded with £9 Prolific credits. The average study duration was 33.6 minutes.

Table 1. First experiment with two counter-balanced groups.

Experiment	Group	Scenario 1	Scenario 2	#
1	A	No Prediction (No-PFM)	Formal Prediction (PFM)	30
	B	Formal Prediction (PFM)	No Prediction (No-PFM)	30

Table 2. Second experiment to further investigate findings in the first experiment by introducing BRPM.

Experiment	Group	Scenario 1	Scenario 2	#
2A	A	No Prediction (No-PFM)	Bounded Random (BRPM)	30
	B	Bounded Random (BRPM)	No Prediction (No-PFM)	30
2B	A	Bounded Random (BRPM)	Formal Prediction (PFM)	30
	B	Formal Prediction (PFM)	Bounded Random (BRPM)	30

Just Accepted

4.1 Simulation Environment

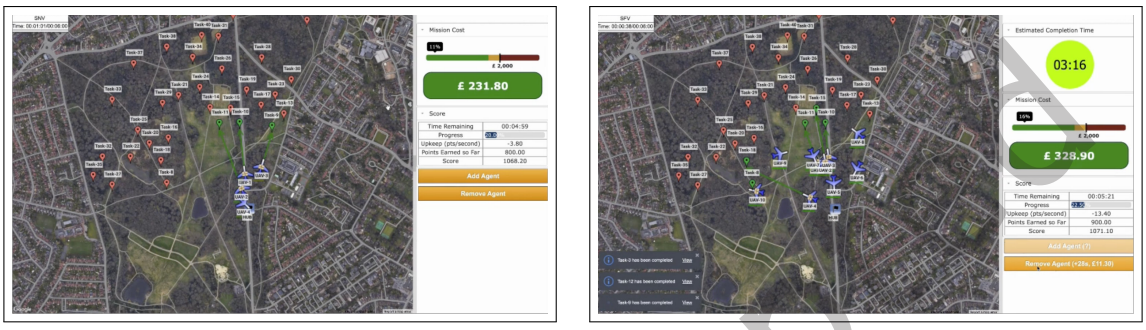
We modelled a swarm of drones and their control system using the HARIS platform. This is a webapp-based simulator which models large numbers of agents in a Google Maps environment. The simulation engine runs on a server which is accessed by the user through a standard web browser. The user's interface is updated and their commands are sent using REST API; the user's view and control are both handled in real-time. Drones in this configuration of the simulator perform a simple behaviour based on centralised best-first task service (any free drone is allocated to the closest task), with tasks being represented by points on the map [25]. This platform allowed quick integration of the modelling as well as logging of the user actions and the ability to perform our study online through the cloud. Each drone's movement is tied to the real-time simulation loop and is based on a simple physics engine, so the velocities, turn speeds, and battery decay are analogous to real robots.

4.2 Scenarios and Interface

To investigate the impact of predictive formal modelling (or its absence), we developed three distinct conditions: PFM, no-PFM, and BRPM. While the task was the same regardless of the condition experienced, the information regarding the completion of the mission which was provided to the swarm operator varied. The No-PFM scenario, acting as a baseline, displayed the interface as presented in Figure 2a. This included the map (left), as well as crucial mission information pertaining the 'mission cost', 'current expenditure', 'time remaining', 'mission progress', 'upkeep cost', 'points earned', and 'score'. Additionally, the interface provided the participant with the possibility to add and remove drones (minimum and maximum active drones: 3–10). Contrasting this, the interface used for the PFM and BRPM condition included information on the 'estimated completion time' as well as the extra cost

[2]Prolfic: https://www.prolific.com
[3]We received ethics approval from the University of Southampton's ethics committee (confirmation number: ERGO/FEPS/85523).

incurred and time gained for adding or removing an extra drone. The estimated completion time was highlighted in green, yellow, or red depending on if the estimated time was below, near, or beyond the six-minute target (see circle in upper right corner, Figure 2b). The estimated cost incurred in pounds (£) and time gained in seconds (s) for adding or removing a drone is displayed in the "add or remove agent" button at the bottom right of the interface. The difference between the PFM and BRPM conditions was the accuracy of the estimated completion time. In the PFM scenario, the participant was provided with an accurate estimation, while the presented completion time in the Bounded Random (BRPM) condition was a misleading but believable value.



(a) No-PFM Interface.

(b) Interface for PFM and BRPM.

Fig. 2. Interfaces used for the three conditions. In contrast to the No-PFM (Figure 2a), the PFM and BRPM (Figure 2b) presented the participant with an estimated completion time (upper right corner) as well as the impact on the price (£) and time (sec) when adding or removing drones.

The formula for the upkeep at time t seconds is given as:

$$U(n_t) = 0.1n_t^2 + 1.3 \tag{4}$$

where n_t is the number of agents at time t seconds. The total mission cost is the sum of the per second upkeep cost for the duration of the mission, and it is given as:

$$\text{Total Cost} = \sum_{t=1}^{T} U(n_t); \quad T \leq 360s \tag{5}$$

where T is the scenario simulation completion time, with $T \leq 360s$ (6 minutes) — the maximum simulation time allowed. The upkeep per-second cost is based on a square of number which is offset by a constant to make it difficult for a participant to predict the cost of adding or removing a UAV, especially for the No-PFM condition without the predictive model. The cost of adding or removing an agent is computed by replacing 'n_t' with '$n_t + 1$' or '$n_t - 1$' in the upkeep formula (Equation 4). This is then displayed on the add or remove button appropriately.

The time gain/loss value shown in the add or remove agent button is calculated using exactly the same procedure in Equation 2 (PFM) and Equation 3 (BRPM), but the state input into the formal model assumes that a drone is added or removed for the add or remove button respectively. This is the same as asking the model "if I add/remove an agent now, and run the model on the new swarm, what will the result be?". Therefore, in addition to computing for the current state, we also compute for two possible future situations, one in which a drone was added and another in which a drone was removed, and display the result appropriately on the add or remove agent button.

Just Accepted

4.3 Study task

The study task in each of the three experiments was based on a drone delivery mission. The participants were required to complete 40 package deliveries within 6 minutes with a budget of £2000 using between 3 – 10 drones. The participants playing the role of the swarm operator managed the mission by adding or removing agents. The more UAVs they added, the faster they completed the mission, but they incurred a higher mission cost. The reverse was also applicable in that the more UAVs they removed from the mission, the longer the mission completion time, and the less the overall mission cost. The final mission cost was a cumulative sum of the upkeep cost per second. We implemented a non-linear per-second upkeep cost function that makes the upkeep per-second cost higher each time a new UAV is added. The non-linearity was chosen to reduce the participant's ability to predict the cost of adding or removing a UAV, especially for the No-PFM condition without the predictive model. It also makes it more costly for the participant to try to make up for their mistakes by adding lots of drones near the end (i.e. it is better to keep at ~7 drones than to keep 5 drones for a while and then increase to 10 drones).

4.4 Procedure

Following recruitment, participants were presented with the participant information sheet and consent form, after which they completed a brief demographics survey which collected data on gender, age group, education level, self-rated computer expertise, as well as self-rated UAV or Swarm robotics knowledge. Subsequently, participants were asked to watch a short study briefing video and asked to answer three questions to test their preliminary understanding of the task. To ensure that participants understood the study task, two of these three had to be answered correctly in order to proceed with the study. Participants were required to perform a short tutorial scenario which allowed them to experiment with all the provided functionality and experience the interface prior to the actual data collection. The tutorial scenario demonstrated the PFM interface. Participants then proceeded to their first scenario. Following its completion, they completed the post-task survey which included the 6-item NASA-TLX [24], the 10-item System Usability Scale (SUS) [11], the 12-item Trust in Automated Systems [26], and the 9-item system acceptance [51] questionnaire[4]. They continued with the second scenario, followed by the same set of questionnaires. Finally, participants were asked to complete a short survey in relation to their preferred scenario condition, before returning to Prolific. These questions were related to (a) the perceived accuracy of the time estimation feature provided (for the PFM and BRPM condition), (b) their preferred scenario, (c) a selection of reasons for perceived success during task completion, (d) the primary reason for their success, and (e) a binary selection if they used the estimated completion time. Each participant's performance was measured and recorded in real-time during the scenario tasks as HARIS generated log files. Finally, a free text input field allowed for final comments.

The survey questionnaires and HARIS simulator were dockerised and deployed online[5] on an AWS EC2 c5.4 x large (32GB RAM, 16 vCPUs) instance running the Ubuntu 22.04 operating system. The dockerisation was necessary for a scalable deployment due to the high computing resource requirement of the prediction model in the simulator. The research dataset, which includes the complete listing of participant survey questions, anonymised participants' questionnaire responses, and their HARIS simulator performance data from each of the three experiments, has been published in [1] for broader public access and availability.

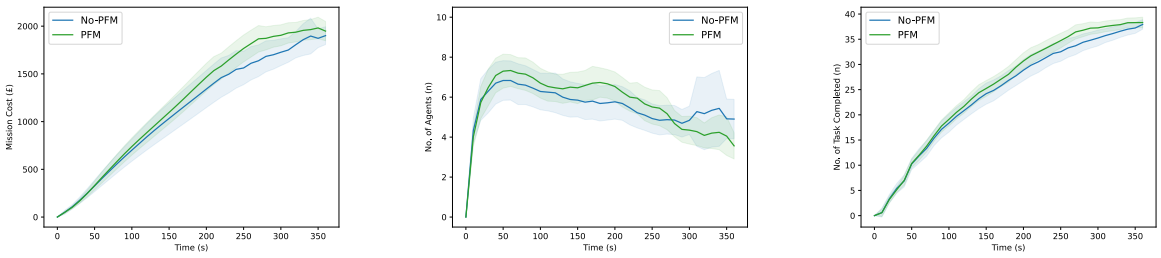
5 RESULTS AND ANALYSIS

This section presents the results from all three experiments. Prior to the data analysis, we confirmed that the dataset of each experiment met the prerequisites necessary for conducting one-way ANOVA testing. Additionally,

[4]Trust and acceptability questionnaire data was not presented or analysed because it failed the Likert scale test for reliability (Cronbach's alpha < 0.7)

[5]Online HARIS simulator: https://uos-haris.online

Just Accepted



(a) Mission Cost (b) Number of Agents (c) Completed Tasks

Fig. 3. Comparing the mean performance over time for Experiment 1. The shaded region around the mean shows the standard deviation.

we performed a G*Power analysis to verify that our sample size of each experiment aligns with the required criteria. Specifically, we have 60 participants in each experiment, with an assumed effect size of 0.2 and a significance level of 0.05. The G*Power analysis helps to confirm that our studies were adequately powered to detect the expected effects.

To compare the impact of the models on the performance of the human-swarm teaming, we evaluated four dependent variables in each of the three experiments. Specifically, we investigated a) the `Time Completion' referring to the mission completion time i.e., the time taken to complete 40 delivery tasks; b) the `No. of Agents' used which refers to the mean number of agents deployed by each participant to complete the delivery task; c) the number of `Completed Tasks', we considered a delivery task to be successfully completed when the UAV reaches the target coordinate before returning to the hub to collect parcels for the next delivery; and d) the `Cost per Task' which was computed as a ratio of the mean total cost incurred over the mean number of tasks completed per study scenario.

5.1 Experiment 1

The result[6] of the participants' performance over time is presented in Figure 3. Figure 3a shows the mean mission cost of each scenario over time. The No-PFM scenario incurred a lower cost over time than the PFM scenario. Figure 3b shows the mean number of agents used over time. Although the PFM group used more agents than the No-PFM group in the first three-quarters of the experiment, the No-PFM group finished with more agents than the PFM group. Since the No-PFM condition did not have the estimated completion time displayed, it is possible that they realised very late that they may not finish, and therefore started adding more agents towards the end. This might indicate that participants in the No-PFM condition found it more difficult to balance the number of agents with the two constraints defined. Figure 3c compares the mean number of completed tasks over time and shows that participants completed more tasks on average over time in their PFM scenario compared to their No-PFM scenario.

Given that the participants were asked to optimise for two objectives – finish within 6 minutes and not exceed a £2000 budget – further analysis was conducted to determine how many participants succeeded at both objectives. Figure 4 shows a plot of the mission cost against completion time for the Experiment 1 participants who completed

[6]The main results in this subsection were previously presented in [2]. Part of this has been repeated for a more comprehensive result presentation and for better readability. In addition, the presentation has been revised and a new time-cost performance and equivalence analysis have been added.

Just Accepted

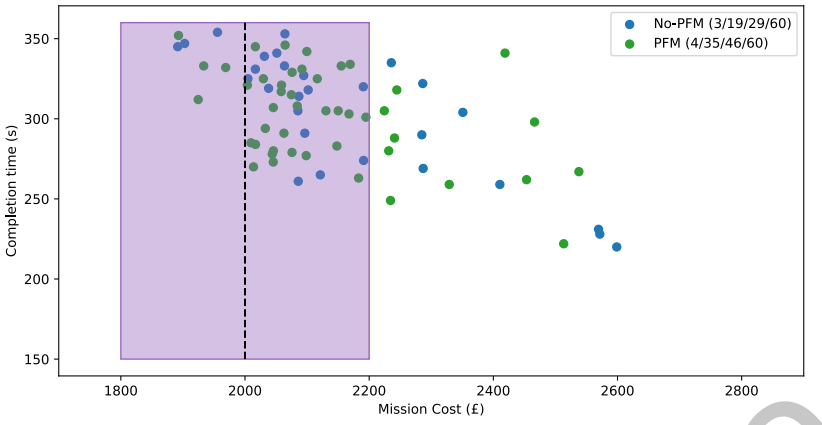


Fig. 4. Plot of mission cost against completion time showing £2000 cost line and its 10% bounded region from £1800 - £2200 for all completed PFM and No-PFM conditions in Experiment 1. The legend shows the number of completions within £2000, the bounded region completions, the total number of completions within time, and the total number of participants in each condition.

the 40 deliveries within the 6-minute mission time[7]. From the figure, only 29 of the 60 (48.3%) participants finished in the No-PFM condition compared to the PFM condition where 46 of the 60 (76.7%) participants finished. However, the number of participants who finished within the £2000 budget was 3 out of 60 (5%) for the No-PFM and 4 out of 60 (6.7%) for the PFM. In other words, the total number of finishes was 7 for the 120 participant scenarios (5.8%). This shows that the budget was too strict and barely attainable. Therefore, an expanded analysis was conducted that considers participants who finished within 10% of the original £2000 budget. The Purple rectangle in Figure 4 shows the 10% bounded completion regions from £1800 - £2200. From the result, 54 of the 120 participant scenarios (45%) finished within the 10% cost completion bound. Of these, 19 of the finishes were for the No-PFM condition and 35 were for the PFM condition. This shows that the PFM model resulted in a lower cost-time performance.

For workload, participants had a mean of 4.77 (SD = 1.50) in the PFM and a mean of 4.74 (SD = 1.56) in the No-PFM scenarios. One-way ANOVA for workload revealed no significant main effect ($F(1, 118) = 0.009$, $p = 0.924$). While our results showed no statistically significant difference in workload between the PFM and No-PFM conditions ($p = 0.924$), this finding does not confirm equivalence. Hence, we used equivalent testing to assess the equivalence of workload between the PFM ($M = 4.77$, $SD = 1.50$) and No-PFM conditions ($M = 4.74$, $SD = 1.56$). The mean workload difference was within the confidence interval of [-0.32, -0.281]. Given our predefined equivalence margin of ±0.5, as established by previous research [34], these results confirm that the PFM and No-PFM conditions are statistically equivalent in terms of mental workload, indicating that the introduction of the PFM does not significantly increase cognitive demand on participants. In line with the guidelines [11, 18], interfaces with a usability testing value of 68 or above are considered good. Mean SUS scores for PFM and No-PFM scenarios were 70.75 (SD = 17.52) and 74.38 (SD = 15.15). This shows that the usability of both systems was good. One-way ANOVA yielded no significant effect on usability ($F(1, 118) = 1.470$, $p = 0.228$). This suggests that the PFM feature did not make the system more or less usable than without it.

[7]Note that the simulator would normally cut-off participants once the 6-minute mission time is reached, the data of any participant not automatically cut-off by the simulator were manually filtered out for the plot in Figure 4.

Just Accepted

Table 3. Descriptive statistics and one-way ANOVA results for Experiment 1. Significance levels: * $p < 0.05$, ** $p < 0.01$

Variable	Scenario	Mean	Std.	F value	p value
Time Completion	No-PFM	329s	36.93	5.363	0.022*
	PFM	314s	34.77		
No. of Agents	No-PFM	5.79	1.06	3.074	0.082
	PFM	6.10	0.86		
Completed Tasks	No-PFM	38.80	1.71	7.255	0.008**
	PFM	39.55	1.30		
Cost per Task	No-PFM	£52.85	7.31	0.001	0.988
	PFM	£52.83	3.80		

As depicted in Table 3, the PFM condition led to enhanced task completion rates and reduced time requirements when compared to the No-PFM scenario where no prediction was presented to participants. Specifically, participants, on average, completed 39.55 tasks (SD No. of Tasks = 1.30) within 314 seconds (SD Time Completion = 34.77) in the PFM condition. This performance contrasted with the No-PFM condition, where participants completed an average of 38.80 tasks (SD No. of Tasks = 1.71) in 329 seconds (SD Time Completion = 36.93). An ANOVA test was conducted and showed that this difference was significant both in terms of Time Completion [F(1,59) = 5.363, p = 0.022*] and No. of Completed Tasks [F(1,59) = 7.255, p = 0.008*]. Moreover, our findings indicate that employing PFM prediction did not influence the utilisation of additional agents or the associated task cost in the context of human-swarm collaboration when compared to scenarios without prediction (No-PFM). In the PFM condition, participants, on average, employed 6.10 agents (SD No. of Agents = 0.86) at an average cost of £52.83 (SD Cost = 3.80) per task. Conversely, in the No-PFM condition, participants used an average of 5.79 agents (SD No. of Agents = 1.06) at a cost of £52.85 (SD Cost = 7.31) per task. An ANOVA test was conducted and showed that this difference between PFM and No-PFM was not significantly different for No. of Agents [F(1,59) = 3.074, p = 0.082] and Cost per task [F(1,59) = 0.988].

The significant results in relation to mission completion times were also reflected in the anecdotal open-ended statements made by participants, we received 20 open-ended statements from the 60 participants (33%) following both conditions[8]. Participants indicated the perceived usefulness of the features provided in the PFM condition as, e.g., expressed by P55:

> *"I found the presence of the estimated completion time feature [PFM] helped me decide whether to add or remove agents, whereas in the first scenario [No-PFM] I was trying to estimate it myself based on the remaining time and the percentage completion of the task."* — P55

indicating the usefulness of the additional information provided to complete the task successfully. A similar sentiment was presented by P45 who describes the use of the PFM feature as a guiding mark for optimising the addition and removal of drones.

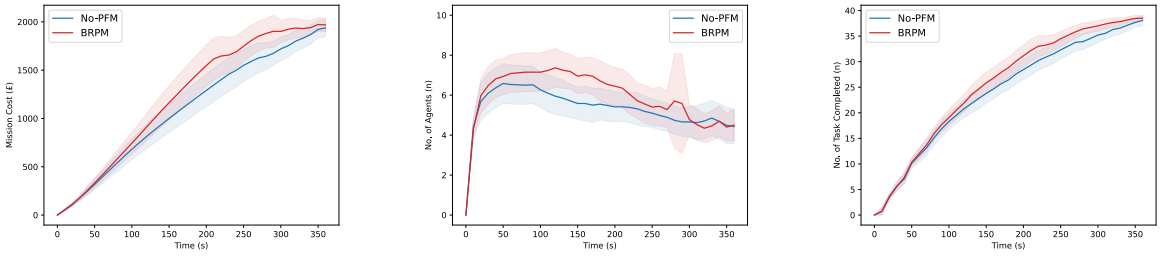
> *"I used the estimated time to allow me to hover around the 6-minute mark, adding and taking away planes where necessary"* — P45

5.2 Experiment 2A

This section presents the outcomes of Experiment 2A and investigates the impact of the Bounded Random (BRPM) condition when compared to the No-PFM baseline. Figure 5 shows the mean performance plot over time for Experiment 2A with the shaded region around the mean representing the standard deviation. Figure 5a shows

[8]The free text input was not required (as described in the Procedure section), and the comments presented below are therefore based on a small number of open-ended responses. These should therefore only be taken as anecdotal and not conclusive evidence.

Just Accepted



(a) Mission Cost (b) Number of Agents (c) Completed Tasks

Fig. 5. Comparing the mean performance over time for Experiment 2A. The shaded region around the mean shows the standard deviation.

that the mean mission cost over time for the No-PFM condition was generally lower than in the BRPM condition. Figure 5b shows the plot of the mean number of agents over time. This showed that more agents were used over time in the BRPM scenario than in the No-PFM scenario. The sharp rise near the 300s mark for the BRPM could be because the participants thought they would fail the mission due to the bounded random prediction suggesting failure even when it was clear they would succeed. This was followed by a sharp fall, perhaps when the participants realised they could not trust the BRPM prediction. Figure 5c shows the mean number of completed tasks over time for the No-PFM and BRPM scenarios. This showed that the BRPM scenario completed more tasks over time than the No-PFM scenario.

Figure 6 shows a plot of the mission cost against completion time for the Experiment 2A participants who completed the 40 deliveries within the 6-minute mission time. From the figure, 25 of the 60 (41.7%) participants finished in the No-PFM condition compared to the BRPM condition where 46 of the 60 (76.7%) participants finished. However, the number of participants who finished within the £2000 budget was 4 out of 60 (6.7%) for the No-PFM condition and 6 out of 60 (10%) for the BRPM condition. In other words, the total number of finishes was 10 for the 120 participant scenarios (8.3%). For the extended bounded completion region (10% of the original £2000 budget extending to £2200), the Purple rectangle region in Figure 6, 44 of the 120 participant scenarios (36.7%) finished within the 10% cost completion bound. Of these, 19 of the finishes were for the No-PFM condition and 25 were for the BRPM condition. This shows that the BRPM model resulted in a lower cost-time performance than the No-PFM model. However, when compared to the PFM in Section 5.1, the PFM model resulted in a 1.4 times lower cost-time performance than the BRPM model, based on the PFM to BRPM completion ratio of 35:25 given the similar No-PFM baseline performance.

For workload, participants had a mean of 5.03 (SD = 1.67) in the BRPM scenario and a mean of 4.78 (SD = 1.44) in the No-PFM scenario. One-way ANOVA for workload revealed no significant main effect ($F(1, 118) = 0.810$, $p = 0.370$. While our results showed no statistically significant difference in workload between the BRPM and No-PFM conditions ($p = 0.370$), this finding does not confirm equivalence. Hence, we used equivalent testing to assess the equivalence of workload between the BRPM (M = 5.03, SD = 1.67) and No-PFM conditions (M = 4.78, SD = 1.44). The mean workload difference was within the confidence interval of [-0.27, -0.33]. Given our predefined equivalence margin of ±0.5, as established by previous research [34], these results confirm that the BRPM and No-PFM conditions are statistically equivalent in terms of mental workload, indicating that the introduction of the BRPM does not significantly increase cognitive demand on participants. Regarding usability, the mean SUS scores for BRPM and No-PFM scenarios were 65.50 (SD = 20.53) and 71.00 (SD = 19.28). Therefore, the BRPM negatively affected the usability of the interface since the SUS score was below the threshold value of 68 whereas

Just Accepted

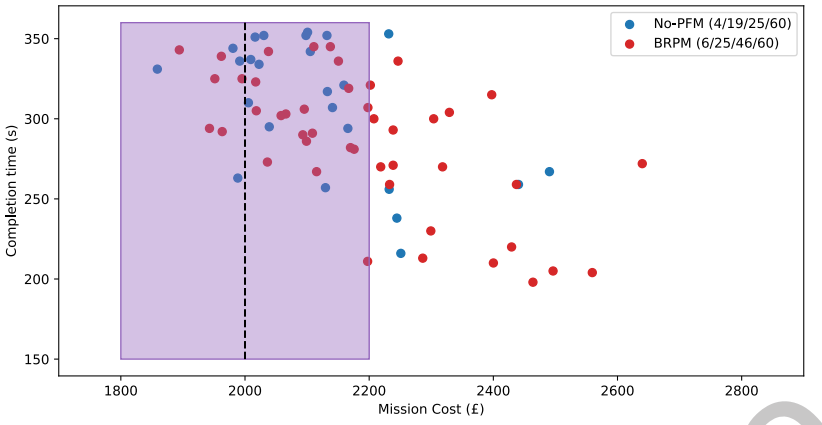


Fig. 6. Plot of mission cost against completion time showing £2000 cost line and its 10% bounded region from £1800 - £2200 for all completed No-PFM and BRPM conditions in Experiment 2A. The legend shows the number of completions within £2000, the bounded region completions, the total number of completions within time, and the total number of participants in each condition.

Just Accepted

the No-PFM was above this value. However, One-way ANOVA yielded no significant effect on usability ($F(1, 118) = 2.288$, $p = 0.133$).

As illustrated in Table 4, the BRPM condition resulted in a significant reduction in task completion time and a noteworthy increase in the overall number of tasks among participants. However, compared to the No-PFM scenario, where no predictions were provided to participants, the BRPM condition led to a significant increase in task cost and the number of agents assigned to each task. On average, participants completed 39.61 tasks (SD No. of Tasks = 0.78) within 302 seconds (SD Time Completion = 46.44) in the BRPM condition. In contrast, in the No-PFM condition, participants completed an average of 38.50 tasks (SD No. of Tasks = 2.04) in 333 seconds (SD Time Completion = 34.08). An ANOVA test indicated that this difference was significant for both Time Completion [$F(1,59) = 17.827$, $p = 0.0001$] and No. of Completed Tasks [$F(1,59) = 15.699$, $p = 0.001$]. Furthermore, our findings reveal that the utilization of the BRPM prediction significantly influences the involvement of additional agents or the associated task cost in the context of human-swarm collaboration when compared to scenarios without prediction (No-PFM). In the BRPM condition, participants, on average, engaged 6.45 agents (SD No. of Agents = 1.23) at an average cost of £54.40 (SD Cost = 5.90) per task. Conversely, in the No-PFM condition, participants used an average of 5.57 agents (SD No. of Agents = 0.85) at a cost of £51.46 (SD Cost = 3.63) per task. ANOVA tests demonstrated significant differences between BRPM and No-PFM for No. of Agents [$F(1,59) = 20.566$, $p = 0.001$] and Cost per task [$F(1,59) = 10.810$, $p = 0.001$].

Finally, when comparing the responses to free-text form to Experiment 1, it becomes apparent that participants in Experiment 2A report a greater frustration towards the estimated completion time presented to them. Both Experiment 1 and 2A experienced the No-PFM condition, making the primary difference the experimental condition experienced: PFM in Experiment 1 and BRPM in Experiment 2A. In Experiment 1 only one of the 20 open-ended statements (5%) commented that "the estimated time to complete was kind of off". In contrast, in Experiment 2A, five participants (25%) of the 20 open-ended statements raised concerns about the perceived inaccuracy of the estimated time presented to them, indicating a lower willingness to rely on the information provided. For instance P80 expressed their frustration by stating that:

Table 4. Descriptive statistics and one-way ANOVA results for Experiment 2A. Significance levels are indicated as *** $p < 0.001$

Variable	Scenario	Mean	Std.	F value	p value
Time Completion	No-PFM	333s	34.08	17.827	0.001***
	BRPM	302s	46.44		
No. of Agents	No-PFM	5.57	0.85	20.566	0.001***
	BRPM	6.45	1.23		
Completed Tasks	No-PFM	38.50	2.04	15.699	0.001***
	BRPM	39.61	0.78		
Cost per Task	No-PFM	£51.46	3.63	10.810	0.001***
	BRPM	£54.40	5.90		

"it made no sense that in the very first scenario [BRPM], the time went from however many mins to zero time left in just a few seconds?!" — P80

5.3 Experiment 2B

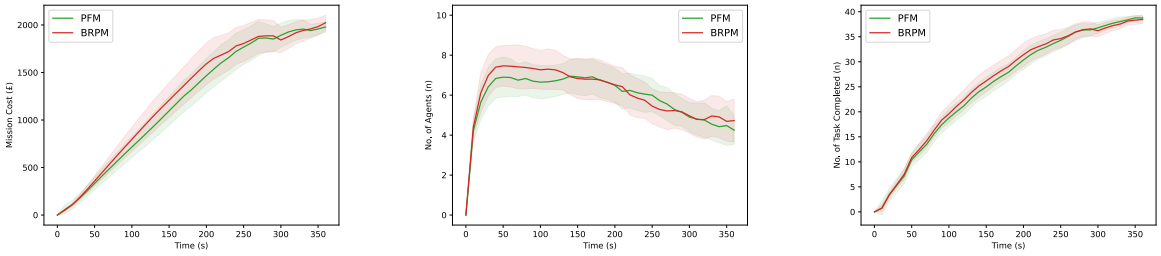
Experiment 2B compared the result of the BRPM condition with the proposed PFM method. By comparing the effects of random and accurate PFM on human-swarm performance, this experiment addresses the specific question of prediction quality. It helps distinguish whether the performance improvements observed in Experiment 1 (accurate prediction) are attributed to the presence of any PFM or specifically to accurate PFM. This comparison provides insights into the importance of PFM reliability in optimizing human-swarm collaboration.

Figure 7 shows the mean performance plot over time for Experiment 2B with the shaded region around the mean representing the standard deviation. Figure 7a shows that the mean mission cost over time for the BRPM condition was slightly higher around the middle of the experiment. Figure 7b shows the plot of the mean number of agents over time. This showed that more agents were used at the beginning by the BRPM scenario than by the PFM scenario. The BRPM scenario also finished with more number of agents than the PFM scenario. Figure 7c shows the mean number of completed tasks over time for the PFM and BRPM scenarios. This showed that the participants completed slightly more tasks around the middle of the experiments in the BRPM scenario than in the PFM scenario.

Figure 8 shows a plot of the mission cost against completion time for the Experiment 2B participants who completed the 40 deliveries within the 6-minute mission time. From the figure, 46 of the 60 (76.7%) participants finished in the PFM condition and 49 of the 60 (81.7%) participants finished in the BRPM condition. However, the number of participants who finished within the £2000 budget was 4 out of 60 (6.7%) for the PFM condition and 3 out of 60 (5%) for the BRPM condition. The total number of finishes within the £2000 budget was 7 for the 120 participant scenarios (5.8%). For the extended bounded completion region (10% of the original £2000 budget extending to £2200), the Purple rectangle region in Figure 8, 56 of the 120 participant scenarios (46.7%) finished within the 10% cost completion bound. Out of these, 30 of the finishes were for the PFM condition and 26 were for the BRPM condition. This shows that the PFM model resulted in a slightly lower cost-time performance than the BRPM model. In other words, the PFM model resulted in a 1.2 times lower cost-time performance than the BRPM model. Although the direct comparison resulted in 1.2 times lower cost-time performance and the indirect comparison (in Section 5.2) showed 1.4 times lower cost-time performance for the PFM over the BRPM, both results agree that PFM generally results in a lower cost-time performance.

For workload, participants had a mean of 5.00 (SD = 1.73) in the BRPM scenario and a mean of 5.19 (SD = 1.52) in the PFM scenario. One-way ANOVA for workload revealed no significant main effect ($F(1, 118) = 0.380$, $p = 0.539$. While our results showed no statistically significant difference in workload between the PFM and No-PFM

Just Accepted



(a) Mission Cost (b) Number of Agents (c) Completed Tasks

Fig. 7. Comparing the mean performance over time for Experiment 2B. The shaded region around the mean shows the standard deviation.

Table 5. Descriptive statistics and one-way ANOVA results for Experiment 2B. Significance levels are indicated as: * $p < 0.05$

Variable	Scenario	Mean	Std.	F value	p value
Time Completion	BRPM	297s	48.42	3.305	0.072
	PFM	313s	43.75		
No. of Agents	BRPM	6.68	1.18	3.960	0.049*
	PFM	6.29	0.99		
Completed Tasks	BRPM	39.53	1.21	0.551	0.459
	PFM	39.67	0.68		
Cost per Task	BRPM	£55.47	4.59	4.591	0.034*
	PFM	£53.51	5.38		

conditions (p = 0.539), this finding does not confirm equivalence. Hence, we used equivalent testing to assess workload equivalence between the BRPM (M = 5.00, SD = 1.73) and PFM conditions (M = 5.19, SD = 1.52). The mean workload difference was within the confidence interval of [-0.37, -0.34]. Given our predefined equivalence margin of ±0.5, as established by previous research [34], these results confirm that the BRPM and PFM conditions are statistically equivalent in terms of mental workload. This suggests that there was no significant change in participants' workload between the BRPM and PFM scenarios. Regarding usability, the mean SUS scores for BRPM and PFM scenarios were 67.21 (SD = 21.22) and 69.33 (SD = 20.98). Therefore, both the BRPM and PFM interface usability were good since both interfaces had SUS scores above the threshold value of 68. One-way ANOVA yielded no significant effect on usability (F(1, 118) = 0.304, p = 0.582). This means there was no significant difference between the two interfaces in terms of their usability.

As shown in Table 5, the Predictive Formal Model (PFM) condition exhibited a noteworthy decrease in cost per task and the number of agents per task compared to the BRPM condition. However, no significant differences were found between time completion and the number of completed tasks when comparing PFM and BRPM conditions.

On average, participants incurred a cost of £55.47 (SD Cost = £4.59) per task in the BRPM condition, whereas they spent £53.51 (SD Cost = £5.38) in the PFM condition. ANOVA tests validated this observation, revealing significant differences between BRPM and PFM [F(1,59) = 4.591, p = 0.034]. The number of agents utilised by participants in each task decreased from 6.68 (SD No. of Agents = 1.18) in the BRPM to 6.29 (SD No. of Agents = 0.99) in the PFM condition. This difference was significant based on the one-way ANOVA test [F(1,59) = 3.960, p = 0.049]. Participants' open-ended statements, 23 of the 60 participants chose to provide a comment (38.3%),

Just Accepted

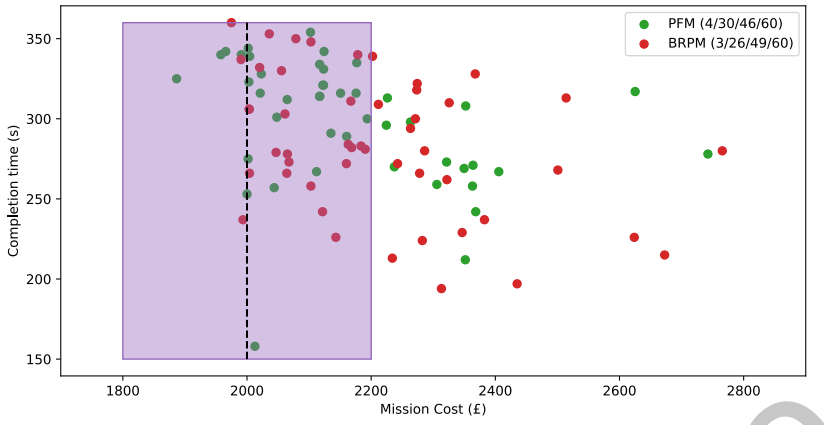


Fig. 8. Plot of mission cost against completion time showing £2000 cost line and its 10% bounded region from £1800 - £2200 for all completed PFM and BRPM conditions in Experiment 2B. The legend shows the number of completions within £2000, the bounded region completions, the total number of completions within time, and the total number of participants in each condition.

Just Accepted

further document the difficulty of balancing the amount of UAVs needed given the inaccuracies of the estimated completion time feature as expressed by P138:

> *"On the first one [PFM], I tried to add and reduce agents often to try and keep it even. On the second one [BRPM] I tried a different approach and didn't add any at first because I had lots of time. However, my estimated time went up faster than I thought this time and I struggled to get it back down. It was a good study though."* — P138

While the findings for the time completion between the two conditions are not significant, participants perceived ability to complete the task within the cost constraint was reported. For instance, P172 reported that they were more capable of completing the task within the time without exceeding the cost in the second attempt (PFM). This aligns with the costs per task being significantly lower (see Table 5).

> *"I went over the cost on my first attempt [BRPM] but felt the estimated time was off, I did better on the second attempt [PFM]..."* — P172

6 DISCUSSION

The major contribution of this work was demonstrating the fact that predictive formal modelling at runtime increased mission performance without affecting the swarm operator's workload through a series of human-swarm interaction experiments. This has implication in resource allocation, planning missions, adaptively responding to changing mission conditions, and managing mission assets.

6.1 Importance and Impact on Predictive Accuracy

Through the integration of predictive formal modelling at runtime, the swarm operator is equipped with an early warning system for critical mission failure. The swarm operator can re-task agents from where they are excess to where they are needed long before a failure becomes impossible to avert. This could mean more lives being saved in a search and rescue mission or more medical and life-saving supplies being delivered to remotely inaccessible

areas in the event of a natural disaster. In the PFM scenarios, agents can be re-tasked without losing time to areas where they are needed unlike in no-PFM scenarios where the operator only gets to know last minute.

In real world applications such as firefighting and disaster relief, the problem of assigning drones becomes much broader. The type of scenario we model might only be one section of a large and continually changing environment. Although we place the constraint of cost on our participants, there may be a higher-level commander who requires the operator to complete this mission with minimal resources so that more drones can be reassigned to other areas. Having a good understanding of expected conditions on the horizon may allow future disaster response teams to distribute their limited resources more efficiently by eliminating redundancy and optimising the most important and necessary tasks.

It is also interesting to observe that inaccurate predictive information can also improve the operator's performance to a certain extent. As shown in Experiment 2A, given a random prediction, participants completed more tasks in a shorter time in the BRPM scenario compared with their performance in the No-PFM scenario. Such observation may indicate that the simplicity of the scenario we considered makes the accurate PFM not needed as operators are capable of assessing the situation with a simple hint that they are going to accomplish the mission in the next few minutes. This is also confirmed by the usability check between the PFM and the No-PFM scenario in Experiment 1 where participants did not find the PFM feature more usable. However, we also noticed that accurate information becomes more needed in severe situations. For example, as the experiment was approaching the finishing point, participants tended to examine the accuracy of the estimated completion time and adjusted the number of agents accordingly. This happened when participants found the BRPM prediction was not accurate and dramatically removed agents. A similar turning point can also be observed in Experiment 1 where participants in the No-PFM scenario realised very late that they might not finish all tasks on time and tried to add more agents. From these observations, future work can focus on implementations in more complex and high-risk scenarios, such as the real-world applications mentioned above, where it is beyond the operator's cognitive ability to understand the current situation. In such cases, we expect the accurate PFM to become more crucial for operators to make appropriate decisions.

There was a clear difference in the reported accuracy of the estimated time feature between the PFM and BRPM conditions when we compared the participants' open-ended comments in Experiment 1 (No-PFM/PFM) and Experiment 2A (No-PFM/BPRM). Numerous participants (~26%) commented on the lack of accuracy of the estimated time in the BRPM, which led to them failing to complete the mission within the given constraints and trusting the system less. The PFM had far fewer negative mentions about the estimated completion time feature (~5%). Furthermore, not only were the PFM participants less sceptical about the time estimation feature, but around (~15%) of the PFM participants in Experiment 1 explicitly highlighted it as a beneficial feature and attributed their success to it.

6.2 Explainability of Algorithmic Decisions

Recent research [21, 35, 39, 50] is increasingly focusing on the value of algorithmic explainability. This becomes increasingly relevant when humans have to rely on algorithmic decisions. Especially when calibrating trust, i.e., preventing overtrust, providing an explanation on why an algorithm makes a specific choice has proven useful.

Leichtmann et al. [35] conducted an online study to examine how participants in high-risk scenarios perform when assisted by an AI system that either provides or withholds explanations for its recommendations. When the AI offered explanations for its decisions, its participants' trust in the system was calibrated, preventing blind overtrust. Since overtrust negatively affected task performance, the calibrated trust resulting from the explanations ultimately improved participants' performance. While the context investigated by [35] (i.e. classification of safe and poisonous mushrooms) is quite different from multi drone management, it stands to reason that the same lesson might apply in this high-risk scenario. In its current form, the research presented in this paper evaluates

Just Accepted

how three different models (No-PFM, PFM, and RPM) perform when compared to each other. However, in addition to the system provided information about time and cost removed/added when changing the number of drones (see Section 4), user performance, and ultimately reliance on the system, might be further improved by providing a rationale, thereby emphasising explainability of the recommendations.

Furthermore, [39] demonstrated that users develop more appropriate levels of trust in algorithmic decision-making when the algorithm provides 'honest' explanations for its decisions. Similarly, a recent study by Suffian [50] not only emphasised the importance of explanations but highlighted the particular value of those when containing actionable information. Relating this literature to the present studies, would could envision a scenario in, e.g., the PFM condition in which the system employs one or both of the following strategies. To emphasise the calibration of overtrust, the system could present the user with information about the confidence in its predictions. For instance, instead of simply displaying "Estimated Completion Time: 03:16" (see Figure 2b), the system could provide a measure of algorithmic confidence, e.g., "Estimated Completion Time: 03:16. Estimate confidence: [Low/Med/High or 25%/50%/75%/X%]". This would allow users to make informed decisions wether the system perceives the current number of deployed drones appropriate, and how confident the system is in this classification.

Taking this a step further, as suggested by Suffian [50], the PFM condition could also offer actionable explanations to help users achieve their goal of parcel delivery within the time and resource limits. For instance, the system could expand on the current suggestion ("Estimated Completion Time: 03:16", see Figure 2b) by adding a recommendation such as "You are currently using more drones than necessary. I suggest reducing the number to minimise mission costs." This approach would aid users in optimising decision-making while calibrating their trust in the system, preventing both over- and undertrust.

6.3 Future Work and Limitations

To simplify the model, this work considered a homogeneous swarm of UAVs; in real-world applications, this is often not the case as different types of agents with differing capabilities are often utilised. Future works could extend the model to consider the case of heterogeneous swarms, for example, a scenario with both aerial and ground robot swarms, each consisting of agents with differing speeds or carrying capacities. This may amplify the utility of PFM, as the mission state becomes harder to comprehend and the outcome more difficult to predict. We also overlooked the confounding effects of adding an additional interface feature to the PFM condition. This being the inclusion of the price increase incurred by adding or removing a drone, which was present on the add and remove agent buttons in the PFM condition but not in the no PFM condition. Another simplification made was to assume that agents would not experience random failure; in the real world drones inevitably crash or lose contact with operators and this should be factored into estimations made by the model in order to increase the ecological validity, and real-world comparability of the studies. An extended model could consider this possibility, as in [23] and ensure that the user tunes the number of drones to this consideration (i.e. having a spare to replace the drone we expect to fail). Failure rates may also change according to location, battery life, age of the drone, or weight of packages and this too could be considered by the model. If the operator was penalised for agent failure then they may be able to avoid this by using the model to anticipate and preserve their UAVs. It also worth noting that the failure of a drone is likely to substantially impact the operator [27], so early warning or reassurance that this is to be expected may provide further benefits to the human user.

Additionally, a key assumption made in this work is the agents have perfect communication with the hub (and each other). Connectivity is a key issue in urban areas, particularly for aerial vehicles. Any perturbation in connectivity would make it impossible to accurately model the swarm and hence reduce the usefulness of PFM. To align with the uncertainty in communication and other uncertainties, such as weather variability and dynamic obstacles, our predictive formal model could be more probabilistic. For example, the status of individual drones, including their failures and successes, would be updated in a probabilistic manner and only confirmed when in

Just Accepted

range of the base station or other drones. In terms of the reasoning process, more advanced properties, not just the overall success rate of the mission, can be considered, for example, multi-objective properties to evaluate the trade-off between the cost and the time limit; safety properties to highlight the drones that are more likely to fail or lose connection; optimisation properties to compute an optimal strategy for operators. To support these, the underlying formal models have to go beyond CTMCs and take the form of, for instance, (Partially Observable) Markov Decision Processes [43], which are typically used in planning under uncertainties. Another assumption we made is that PFM predictions calculated from SMC approximations are accurate and will not bring substantial impact on users' performance. Because the add or remove button relies on the model's accuracy, this may slightly impact the accuracy of the time prediction suggested to the user for adding or removing a UAV. A future study could provide an option of using the exact model checking approach allowing users to decide whether to wait for a longer period to obtain the exact prediction. The user study design focused on the recruitment of participants who were not experienced swarm operators and had to be trained to use the HARIS platform. A future study could focus on how the performance of expert swarm operators, e.g. with at least 5 years of experience, is affected by predictive formal modelling at runtime. Also, software agents were used in the HARIS simulation in this study. Digital twins of real UAVs can also be integrated into the HARIS platform, so in a future work we could investigate how the swarm operator's response changes when their decisions have a real-world impact, or when they are comparing the PFM with robots they can physically see. This would also have the benefit of making our system ready to be deployed in practical real-world applications.

This research focused on evaluating the model driving the time estimation feature (PFM or BRPM) via a user study, however, there is an opportunity for further research on (a) other types of prediction models and (b) how the time estimation is presented to the operator (user interface design). In the case of the model, how does poor estimation affect the operator's trust in the system? Can trust be recovered if lost due to wrong prediction? In terms of the user interface, how would different ways of expressing uncertainty about completion times affect the usability or workload? Could the time estimation be presented in a way that is explainable to the operator, for example, an explanation of how the time estimate was constructed or the level of uncertainty of the prediction? Another future direction is to explore the impact of PFM on the performance, behaviour, or coordination of a decentralised multi-agent system or robot swarm.

7 CONCLUSION

In this paper, we built on previous works in human-swarm interaction by deploying predictive formal modelling (PFM) at runtime and conducted two user studies to determine the effect on usability, workload, and performance. We recruited 180 participants to perform the role of an aerial swarm operator facilitating the delivery of parcels to target locations in a simulation environment. The role required the participants to add or remove agents as needed to complete the mission within the given time and budget. We showed that predictive formal modelling at runtime increased the performance of the human swarm team without affecting the operators' workload or the systems' usability. We discussed the implication of this in resource allocation and managing mission assets where the swarm operator could respond to the early warning for critical mission failure by re-tasking agents from where they are excess to where they are needed long before a failure becomes impossible to avert. Finally, we proposed areas for future research which included modelling random agent failures to determine their effect on the human-swarm interaction, exploring PFM deployment in heterogeneous swarms, investigating the impact of PFM on multi-agent coordination, and field-testing in a real-world environment.

ACKNOWLEDGMENTS

This project was done as part of the EPSRC Smart Solutions Towards Cellular-Connected Unmanned Aerial Vehicles System (EP/W004364/1) and Explainable Human-swarm Systems project funded by UKRI Trustworthy

Just Accepted

Autonomous Systems Hub (EP/V00784X/1). Some of the researchers in this work were supported by the UKRI MINDS-CDT grant (EP/S024298/1) and an Amazon Research Award on Automated Reasoning.

REFERENCES

[1] Ayodeji O. Abioye, William Hunt, Yue Gu, Eike Schneiders, Mohammad Naiseh, Blair Archibald, Michele Sevegnani, Sarvapali D. Ramchurn, Joel E. Fischer, and Mohammad D. Soorati. 2025. Research Data for Paper: A User Study Evaluation of Predictive Formal Modelling at Runtime in Human-Swarm Interaction. (3 2025). doi:10.6084/m9.figshare.28687238

[2] Ayodeji O. Abioye, William Hunt, Yue Gu, Eike Schneiders, Mohammad Naiseh, Joel E. Fischer, Sarvapali D. Ramchurn, Mohammad D. Soorati, Blair Archibald, and Michele Sevegnani. 2024. The Effect of Predictive Formal Modelling at Runtime on Performance in Human-Swarm Interaction. In *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction (HRI '24)*. Association for Computing Machinery, New York, NY, USA, 172–176. doi:10.1145/3610978.3640725

[3] Ayodeji O. Abioye, Mohammad Naiseh, William Hunt, Jediah Clark, Sarvapali D. Ramchurn, and Mohammad D. Soorati. 2023. The Effect of Data Visualisation Quality and Task Density on Human-Swarm Interaction. In *2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. 1494–1501. doi:10.1109/RO-MAN57019.2023.10309454

[4] Christel Baier, Boudewijn Haverkort, Holger Hermanns, and J-P Katoen. 2003. Model-checking algorithms for continuous-time Markov chains. *IEEE Transactions on software engineering* 29, 6 (2003), 524–541.

[5] Ezio Bartocci, Yliès Falcone, Adrian Francalanza, and Giles Reger. 2018. Introduction to runtime verification. *Lectures on Runtime Verification: Introductory and Advanced Topics* (2018), 1–33.

[6] Taha Benarbia and Kyandoghere Kyamakya. 2021. A literature review of drone-based package delivery logistics systems and their implementation feasibility. *Sustainability* 14, 1 (2021), 360.

[7] Nelly Bencomo, Svein Hallsteinsen, and Eduardo Santana De Almeida. 2012. A view of the dynamic software product line landscape. *Computer* 45, 10 (2012), 36–41.

[8] Gordon Blair, Nelly Bencomo, and Robert B France. 2009. Models@ run. time. *Computer* 42, 10 (2009), 22–27.

[9] Serena Booth, James Tompkin, Hanspeter Pfister, Jim Waldo, Krzysztof Gajos, and Radhika Nagpal. 2017. Piggybacking robots: Human-robot overtrust in university dormitory security. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*. 426–434.

[10] Ioana Boureanu, Panagiotis Kouvaros, and Alessio Lomuscio. 2016. Verifying security properties in unbounded multiagent systems. In *Proceedings of the 2016 international conference on autonomous agents & multiagent systems*. 1209–1217.

[11] John Brooke. 1995. SUS: A quick and dirty usability scale. *Usability Eval. Ind.* 189 (11 1995).

[12] Daniel S. Brown, Michael A. Goodrich, Shin-Young Jung, and Sean Kerman. 2016. Two invariants of human-swarm interaction. *J. Hum.-Robot Interact.* 5, 1 (mar 2016), 1–31. doi:10.5898/JHRI.5.1.Brown

[13] Muffy Calder and Michele Sevegnani. 2014. Modelling IEEE 802.11 CSMA/CA RTS/CTS with stochastic bigraphs with sharing. *Formal Aspects of Computing* 26 (2014), 537–561.

[14] Muffy Calder and Michele Sevegnani. 2019. Stochastic Model Checking for Predicting Component Failures and Service Availability. *IEEE Transactions on Dependable and Secure Computing* 16, 1 (2019), 174–187. doi:10.1109/TDSC.2017.2650901

[15] Zhe Chen, Javier Alonso-Mora, Xiaoshan Bai, Daniel D Harabor, and Peter J Stuckey. 2021. Integrated task assignment and path planning for capacitated multi-agent pickup and delivery. *IEEE Robotics and Automation Letters* 6, 3 (2021), 5816–5823.

[16] Jediah R Clark, Mohammad Naiseh, Joel Fischer, Marisé Galvez Trigo, Katie Parnell, Mario Brito, Adrian Bodenmann, Sarvapali D Ramchurn, and Mohammad D Soorati. 2022. Industry Led Use-Case Development for Human-Swarm Operations. In *AAAI 2022 Spring Symposium Series (Putting AI in the Critical Loop: Assured Trust and Autonomy in Human-Machine Teams)*. AAAI, 1–6.

[17] Jason R Cody, Karina A Roundtree, and Julie A Adams. 2021. Human-collective collaborative target selection. *ACM Transactions on Human-Robot Interaction (THRI)* 10, 2 (2021), 1–29.

[18] DHHS. 2013. System Usability Scale (SUS) - Department of Health and Human Services. https://www.usability.gov/how-to-and-tools/methods/system-usability-scale.html

[19] Agner Krarup Erlang. 1917. Solution of some problems in the theory of probabilities of significance in automatic telephone exchanges. *Post Office Electrical Engineer's Journal* 10 (1917), 189–197.

[20] Yliès Falcone, Srđan Krstić, Giles Reger, and Dmitriy Traytel. 2021. A taxonomy for classifying runtime verification tools. *International Journal on Software Tools for Technology Transfer* 23, 2 (2021), 255–284.

[21] Andrea Ferrario and Michele Loi. 2022. How Explainability Contributes to Trust in AI. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 1457–1466. doi:10.1145/3531146.3533202

[22] Xin Gong, Tieniu Wang, Tingwen Huang, and Yukang Cui. 2022. Toward Safe and Efficient Human–Swarm Collaboration: A Hierarchical Multi-Agent Pickup and Delivery Framework. *IEEE Transactions on Intelligent Vehicles* 8, 2 (2022), 1664–1675.

Just Accepted

[23] Yue Gu, William Hunt, Blair Archibald, Mengwei Xu, Michele Sevegnani, and Mohammad D. Soorati. 2023. Successful Swarms: Operator Situational Awareness with Modelling and Verification at Runtime. In *2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. 541–548. doi:10.1109/RO-MAN57019.2023.10309626
[24] Sandra G Hart. 1986. NASA task load index (TLX). (1986).
[25] William Hunt, Jack Ryan, Ayodeji O. Abioye, Sarvapali D. Ramchurn, and Mohammad D Soorati. 2023. Demonstrating Performance Benefits of Human-Swarm Teaming. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems* (London, United Kingdom). IFAAMAS, Richland, SC, 3062–3064.
[26] Jiun-Yin Jian, Ann M Bisantz, and Colin G Drury. 2000. Foundations for an empirically determined scale of trust in automated systems. *International journal of cognitive ergonomics* 4, 1 (2000), 53–71.
[27] Zahra Rezaei Khavas, Amin Majdi, S Reza Ahmadzadeh, and Paul Robinette. 2023. Human trust after drone failure: Study of the effects of drone type and failure type on human-drone trust. In *2023 20th International Conference on Ubiquitous Robots (UR)*. IEEE, 685–692.
[28] Bing Cai Kok and Harold Soh. 2020. Trust in robots: Challenges and opportunities. *Current Robotics Reports* 1 (2020), 297–309.
[29] Andreas Kolling, Katia Sycara, Steve Nunnally, and Michael Lewis. 2013. Human swarm interaction: An experimental study of two types of interaction with foraging swarms. *Journal of Human-Robot Interaction* 2, 2 (2013).
[30] Andreas Kolling, Phillip Walker, Nilanjan Chakraborty, Katia Sycara, and Michael Lewis. 2015. Human interaction with robot swarms: A survey. *IEEE Transactions on Human-Machine Systems* 46, 1 (2015), 9–26.
[31] Thomas Kosch, Robin Welsch, Lewis Chuang, and Albrecht Schmidt. 2023. The placebo effect of artificial intelligence in human–computer interaction. *ACM Transactions on Computer-Human Interaction* 29, 6 (2023), 1–32.
[32] Panagiotis Kouvaros and Alessio Lomuscio. 2015. Verifying emergent properties of swarms. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*.
[33] Marta Kwiatkowska, Gethin Norman, and David Parker. 2011. PRISM 4.0: Verification of probabilistic real-time systems. In *Computer Aided Verification: 23rd International Conference, CAV 2011, Snowbird, UT, USA, July 14-20, 2011. Proceedings 23*. Springer, 585–591.
[34] Daniël Lakens, Anne M. Scheel, and Peder M. Isager. 2018. Equivalence Testing for Psychological Research: A Tutorial. *Advances in Methods and Practices in Psychological Science* 1, 2 (2018), 259–269. doi:10.1177/2515245918770963
[35] Benedikt Leichtmann, Christina Humer, Andreas Hinterreiter, Marc Streit, and Martina Mara. 2023. Effects of Explainable Artificial Intelligence on trust and human behavior in a high-risk decision task. *Computers in Human Behavior* 139 (2023), 107539. doi:10.1016/j.chb.2022.107539
[36] Minghua Liu, Hang Ma, Jiaoyang Li, and Sven Koenig. 2019. Task and path planning for multi-agent pickup and delivery. In *Proceedings of the International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS)*.
[37] Alessio Lomuscio and Edoardo Pirovano. 2021. Verifying fault-tolerance in probabilistic swarm systems. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*. 325–331.
[38] Hang Ma, Craig Tovey, Guni Sharon, TK Kumar, and Sven Koenig. 2016. Multi-agent path finding with payload transfers and the package-exchange robot-routing problem. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 30.
[39] Siddharth Mehrotra, Carolina Centeio Jorge, Catholijn M. Jonker, and Myrthe L. Tielman. 2024. Integrity-based Explanations for Fostering Appropriate Trust in AI Agents. *ACM Trans. Interact. Intell. Syst.* 14, 1, Article 4 (Jan. 2024), 36 pages. doi:10.1145/3610578
[40] Amirhossein H. Memar and Ehsan T. Esfahani. 2019. Objective Assessment of Human Workload in Physical Human-robot Cooperation Using Brain Monitoring. *J. Hum.-Robot Interact.* 9, 2, Article 13 (dec 2019), 21 pages. doi:10.1145/3368854
[41] Mohammad Naiseh, Dena Al-Thani, Nan Jiang, and Raian Ali. 2021. Explainable recommendation: when design meets trust calibration. *World Wide Web* 24, 5 (2021), 1857–1884.
[42] Mohammad Naiseh, Mohammad D Soorati, and Sarvapali Ramchurn. 2023. Outlining the design space of eXplainable swarm (xSwarm): experts perspective. *arXiv preprint arXiv:2309.01269* (2023).
[43] Gethin Norman, David Parker, and Xueyi Zou. 2017. Verification and control of partially observable probabilistic systems. *Real-Time Systems* 53 (2017), 354–402.
[44] Katie J Parnell, Joel E Fischer, Jediah R Clark, Adrian Bodenmann, Maria Jose Galvez Trigo, Mario P Brito, Mohammad D Soorati, Katherine L Plant, and Sarvapali D Ramchurn. 2022. Trustworthy UAV relationships: Applying the Schema Action World taxonomy to UAVs and UAV swarm operations. *International Journal of Human–Computer Interaction* (2022), 1–17.
[45] Sarvapali D Ramchurn, Joel E Fischer, Yuki Ikuno, Feng Wu, Jack Flann, and Antony Waldock. 2015. A study of human-agent collaboration for multi-UAV task allocation in dynamic environments. In *Proceedings of the 24th International Joint Conference on Artificial Intelligence (IJCAI)*. 1184–1192.
[46] Paul Robinette, Wenchen Li, Robert Allen, Ayanna M Howard, and Alan R Wagner. 2016. Overtrust of robots in emergency evacuation scenarios. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 101–108.
[47] Oren Salzman and Roni Stern. 2020. Research challenges and opportunities in multi-agent path finding and multi-agent pickup and delivery problems. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*. 1711–1715.
[48] Eike Schneiders, EunJeong Cheon, Jesper Kjeldskov, Matthias Rehm, and Mikael B. Skov. 2022. Non-Dyadic Interaction: A Literature Review of 15 Years of Human-Robot Interaction Conference Publications. *J. Hum.-Robot Interact.* 11, 2, Article 13 (feb 2022), 32 pages.

Just Accepted

doi:10.1145/3488242

[49] Mohammad D. Soorati, Mohammad Naiseh, William Hunt, Katie Parnell, Jediah Clark, and Sarvapali D. Ramchurn. 2024. 7 - Enabling trustworthiness in human-swarm systems through a digital twin. In *Putting AI in the Critical Loop*, Prithviraj Dasgupta, James Llinas, Tony Gillespie, Scott Fouse, William Lawless, Ranjeev Mittu, and Donald Sofge (Eds.). Academic Press, 93–125. doi:10.1016/B978-0-443-15988-6.00008-X

[50] Muhammad Suffian. 2023. Explainable AI Assisted Decision-Making and Human Behaviour. In *International Conference on Computing, Intelligence and Data Analytics*. Springer, 376–385.

[51] Jinke D Van Der Laan, Adriaan Heino, and Dick De Waard. 1997. A simple procedure for the assessment of acceptance of advanced transport telematics. *Transportation Research Part C: Emerging Technologies* 5, 1 (1997), 1–10.

[52] Steeven Villa, Robin Welsch, Alena Denisova, and Thomas Kosch. 2024. Evaluating Interactive AI: Understanding and Controlling Placebo Effects in Human-AI Interaction. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–4.

[53] Phillip Walker, Saman Amirpour Amraii, Michael Lewis, Nilanjan Chakraborty, and Katia Sycara. 2014. Control of swarms with multiple leader agents. In *2014 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, 3567–3572.

[54] Håkan LS Younes and Reid G Simmons. 2002. Probabilistic verification of discrete event systems using acceptance sampling. In *Computer Aided Verification: 14th International Conference, CAV 2002 Copenhagen, Denmark, July 27–31, 2002 Proceedings 14*. Springer, 223–235.

[55] Tian Zhou, Jackie S. Cha, Glebys Gonzalez, Juan P. Wachs, Chandru P. Sundaram, and Denny Yu. 2020. Multimodal Physiological Signals for Workload Prediction in Robot-assisted Surgery. *J. Hum.-Robot Interact.* 9, 2, Article 12 (jan 2020), 26 pages. doi:10.1145/3368589

Received 6 May 2024; revised 14 March 2025; accepted 28 March 2025

Just Accepted

University of Glasgow

Abioye, A. O., Hunt, W., Gu, Y., Schneiders, E., Naiseh, M., Archibald, B., Sevegnani, M., Ramchurn, S. D., Fishcer, J. E. and Soorati, M. (2025) A user study evaluation of predictive formal modelling at runtime in human-swarm interaction. *ACM Transactions on Human-Robot Interaction*, (doi: 10.1145/3727989)

There may be differences between this version and the published version. You are advised to consult the publisher's version if you wish to cite from it.

© 2025 Copyright held by the owner/author(s). This is the author's version of the work. It is posted here for your personal use. Not for redistribution. The definitive Version of Record was published in *ACM Transactions on Computer-Human Interaction*, https://doi.org/10.1145/3727989

http://eprints.gla.ac.uk/352658/

Deposited on: 07 April 2025

Enlighten – Research publications by members of the University of Glasgow
http://eprints.gla.ac.uk