The Effect of Predictive Formal Modelling at Runtime on Performance in Human-Swarm Interaction

Ayodeji O. Abioye
Electronics and Computer Science,
University of Southampton,
Southampton, United Kingdom
a.o.abioye@soton.ac.uk

William Hunt
Electronics and Computer Science,
University of Southampton,
Southampton, United Kingdom
w.hunt@soton.ac.uk

Yue Gu
School of Computing Science,
University of Glasgow,
Glasgow, United Kingdom
yue.gu@glasgow.ac.uk

Eike Schneiders
School of Computer Science,
University of Nottingham,
Nottingham, United Kingdom
eike.schneiders@nottingham.ac.uk

Mohammad Naiseh
Computing and Informatics,
Bournemouth University,
Bournemouth, United Kingdom
mnaiseh1@bournemouth.ac.uk

Joel E. Fischer
School of Computer Science,
University of Nottingham,
Nottingham, United Kingdom
joel.fischer@nottingham.ac.uk

Sarvapali D. Ramchurn
Electronics and Computer Science,
University of Southampton,
Southampton, United Kingdom
sdr1@soton.ac.uk

Mohammad D. Soorati
Electronics and Computer Science,
University of Southampton,
Southampton, United Kingdom
m.soorati@soton.ac.uk

Blair Archibald
School of Computing Science,
University of Glasgow,
Glasgow, United Kingdom
blair.archibald@glasgow.ac.uk

Michele Sevegnani
School of Computing Science,
University of Glasgow,
Glasgow, United Kingdom
michele.sevegnani@glasgow.ac.uk

ABSTRACT

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. In this paper, we use predictive formal modelling (PFM) at runtime in a human-swarm mission and show that this integration can be used to improve the performance of human-swarm teams. We recruited 60 participants to operate a simulated aerial swarm to deliver parcels to target locations. In the PFM condition, operators were informed of the estimated completion times given the number of drones deployed, whereas, in the No-PFM condition, operators did not have this information. The operators could control the mission by adding or removing drones from the mission and thereby, increasing or decreasing the overall mission cost. The evaluation of human-swarm performance relied on four metrics: the task completion time, the number of agents, the number of completed tasks, and the cost per task. Our results show that PFM modelling at runtime improves mission performance without significantly affecting the operator's workload or the system's usability.

CCS CONCEPTS

• **Computing methodologies** → **Mobile agents**; **Model verification and validation**; • **Human-centered computing** → **Empirical studies in interaction design**.

KEYWORDS

Human-Robot Interaction (HRI), Human-Swarm Interaction (HSI), Predictive Formal Modelling (PFM), Task Performance

1 INTRODUCTION

Aerial swarms amplify our ability to observe and engage with areas that are challenging for us to reach or oversee. One of the promising applications of aerial swarms is in search and rescue missions to locate and identify casualties on time or to deliver essential life-saving supplies to remote and difficult-to-reach areas in the aftermath of a natural disaster [1]. Such applications are often accompanied by several challenges, ranging from design and deployment [14] to issues related to safety [17], regulations [19], and the operator's mental workload [2]. Other challenges include performance, shared

control and degree of automation, as well as determining the appropriate timing for presenting necessary information to the human operator. Prior research on human-swarm interaction (HSI) has identified essential prerequisites for the successful operation of aerial swarms [9]. 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. Gu et al. [12] 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. This capability allows the swarm operator to make informed decisions promptly, increasing the overall efficiency and adaptability of the human-swarm team.

In this paper, we integrate PFM into the 'Human And Robot Interactive Swarm' (HARIS) simulator [15] to provide human swarm operators with real-time mission and swarm status updates, along with predictions of mission success. Following a within-subject design, we recruited 60 participants to complete a human-swarm task of delivering packages to different areas with two conditions (with and without PFM). We assess the impact of PFM on the performance and required mental workload during the completion of a task as a human-swarm team. Results of the study showed that the PFM condition was able to significantly improve mission performance without having a significant effect on the operator's workload.

2 RELATED WORK

In Abioye et al. [1], the authors evaluated the effect of adding an extra feature to the human-swarm interface, i.e. operator option to request high-quality images of a search area. They found that this led to higher trust perception but did not enhance the overall human-swarm performance. Schneiders et al. [24] indicated the demand for studying non-dyadic human-in-the-loop system configurations, such as that presented in this work. Hunt et al. [15] proposed a method of dynamic re-tasking and triage based on operator feedback. Wilson et al. [27] identified key challenges for the deployment and use of robot swarms, which included how humans understand, monitor, control, and interact with swarms.

Kouvaros and Lomuscio [16] show a formal model for swarms and determine whether the emergent behaviour is satisfied; Boureanu et al. [6] analyse unbounded swarm systems with respect to security requirements by verifying a parameterised model; Lomuscio and Pirovano [18] outline a verification procedure to reason about the fault-tolerance in probabilistic swarm systems. However, none of these approaches can give guarantees after deployment. A close approach to ours is runtime monitoring [3], where pre-constructed monitors are used to analyse the system execution traces that are generated at runtime against formal specifications [11]. These monitors can evolve with system dynamics, such as the size and topology, but cannot reason about mission-level specifications, like the human-swarm interactions, where finite observations are not sufficient. Instead, Gu et al. [12] propose a framework to integrate runtime modelling [5], that has been deployed in system reasoning for unforeseen situations during execution [4], 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 [8].
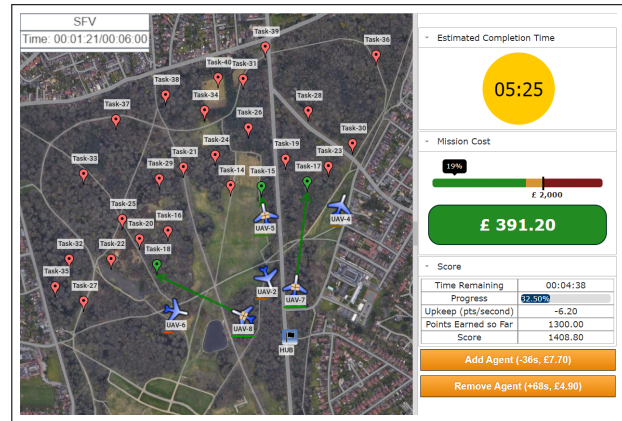

Figure 1: HARIS simulator interface showing the predictive formal modelling based estimated completion time feature in the PFM scenario.

In this work, we adopt an existing model [12] implementing PFM at runtime, to predict the feasibility of a human-swarm mission succeeding and check whether presenting PFM output in the user interface can support human operators in their decision-making during the human-swarm task.

3 STUDY

We conducted a within-subject user study with 60 participants divided into two counterbalanced groups as shown in Table 1, in order to directly compare usability, workload, and performance between the PFM and No-PFM conditions.

Table 1: Showing counterbalanced distribution of the 60 recruited participants.

Group	Scenario 1	Scenario 2	#
1	No Prediction (No-PFM)	With Prediction (PFM)	30
2	With Prediction (PFM)	No Prediction (No-PFM)	30

3.1 Participants

We recruited 60 participants (38 female, 22 male, average age: 34.6, age range 18 - 64) through Prolific [22][1]. 70% of participants have at least a bachelor's degree, 50% were above-average computer users, and 47% were familiar with UAV or swarm robotics. Participants were recruited from the US and the UK. Participants were randomly allocated into groups and received £9. The average study duration was 33.6 minutes.

3.2 Study task

To investigate the impact of PFM, in a human-swarm interaction context, we developed a drone delivery mission scenario presented through two scenarios: the PFM and No-PFM scenarios. To constrain the operators' strategies, we added a cost (£2,000) and time (6 minutes) limit. Users could control the swarm by performing two operations: adding or removing UAVs from the mission. The more

---

[1]Ethics approval from the University of Southampton (ERGO/FEPS/85523).

UAVs they add, the faster they complete the mission, but they incur a higher mission cost (and vice-versa). We set the minimum and maximum number of UAVs allowed in the mission to be 4 and 10 respectively. 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. This reduces the participant's ability to predict the cost of adding or removing a UAV, especially for the No-PFM condition without the predictive model. We added 40 delivery tasks to each scenario. PFM was used to predict the probability of completing all deliveries with the given number of UAVs. This prediction was presented to the operator as an estimated completion time, see upper right corner (see Figure 1). The colour of the circle changes from green to yellow and red depending on the estimated time of completion.

3.3 HARIS Model

The HARIS simulator is a browser-based platform designed for human-in-the-loop multi-agent and swarm robotics experimentation[2]. HARIS is a successor of HutSim [23] which was designed with a focus on usability by consulting with industry experts to model not only their typical command structure but also make it operable with real-life or simulated agents. Building on its predecessor, HARIS was further tailored to its use case derived from interviews with drone pilots [21] and swarm experts [20] to make the platform as usable and realistic as possible while maximising the ease of use for multiple human operators [25], making this simulator a useful tool for the investigations on human-swarm simulations.

We use the model from [12] with slight modifications to better reflect the scenario, e.g., slight changes to the background failure rate of each region, and using the Erlang-k law [10] to provide a smooth transition delay in the continuous-time Markov chain (CTMC). The integration with the HARIS simulator follows a similar process to [12], but with the Sim2PRISM middleware directly embedded in HARIS. Instead of showing the probability of success directly, which might be difficult for participants to interpret, we consider the feasibility over different time intervals and give an estimated completion time as the time when the probability of success reaches 0.99. Additionally, we implement What-if scenarios to give the participants extra information on the effect of adding/removing a UAV before making a decision.

3.4 Procedure

Following recruitment, participants were presented with the participant information and consent form, which was followed by a brief demographics survey collecting data on gender, age group, education, and self-rated computer expertise UAV or Swarm robotics knowledge. Subsequently, participants were asked to watch a short study briefing video and 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. Participants performed a short tutorial which allowed them to experience the simulator's functionality and interface prior to the actual data collection. Participants then proceeded to their first scenario. Following its completion, they

completed the post-task survey which included the 6-item NASA-TLX [13] questionnaire and the 10-item System Usability Scale (SUS) [7]. They continued with the second scenario, followed by the same set of questionnaires. Finally, participants completed a short survey about their preferred scenario, before returning to Prolific. These questions were related to (a) the perceived accuracy of the time estimation feature (in the PFM condition), (b) their preferred scenario, (c) 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.

4 RESULTS AND ANALYSIS

For workload, participants had a mean of 4.77 (SD = 1.50) in the PFM and 4.74 (SD = 1.56) in the No-PFM scenarios. One-way ANOVA for workload revealed no significant main effect ($F(1, 118) = 0.009$, $p = 0.924$). This suggests that the PFM feature did not add additional workload. Regarding usability, we used the system usability scale (SUS) to compare the mean values for the two conditions [7, 26].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 affect the system's usability.

The result of the participants' performance over time is presented in Figure 2. The No-PFM scenario incurred a lower cost over time than the PFM scenario (Figure 2a). The No-PFM group started with the least number of agents but finished with the most (Figure 2b). The No-PFM group did not have the estimated completion time displayed, they might have realised very late that they may not finish, and added 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. Participants completed more tasks on average over time in their PFM scenario compared to their No-PFM scenario (Figure 2c).

We evaluated four dependent variables: a) `Time Completion`: refers to the mission completion time i.e. the time taken to complete 40 delivery tasks. b) `No. of Agents`: refers to the mean number of agents deployed by each participant to complete the delivery task. c) `Completed Tasks`: We considered a delivery task to be successfully completed when the UAV reaches the target coordinate. After this, the UAV returns to the hub to collect parcels for the next delivery. d) `Cost per Task`: This was computed as a ratio of the mean total cost incurred over the mean number of tasks completed per scenario.

Table 2: Descriptive statistics and one-way ANOVA results. Significance levels are indicated as: * $p < 0.05$, ** $p < 0.01$

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		

---

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

(a) Mission Cost

(b) Number of Agents

(c) Completed Tasks

Figure 2: Comparing the mean performance over time. The shaded region around the mean shows the standard deviation.

Our dataset meets the prerequisites necessary for conducting one-way ANOVA testing. We performed a G*Power analysis to verify that our sample size aligns with the required criteria. With $N = 60$ across experimental groups, an assumed effect size of 0.2, and a significance level of 0.05. The G*Power analysis confirmed that our study is adequately powered to detect the expected effects.

As depicted in Table 2, the PFM condition led to enhanced task completion rates and reduced time requirements when compared to the No-PFM scenario. 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 [$F_{(1,59)}= 0.988$]. Figure 2a shows the mean mission cost of each scenario over time.

The significant results in relation to mission completion times were also reflected in the anecdotal open-ended statements. Participants indicated the perceived usefulness of the features provided in the PFM condition as, e.g., expressed by P46: *"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."* (P46) indicating the usefulness of the additional information provided to complete the task successfully. A similar sentiment was presented by P20 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" (P20).

5 DISCUSSION AND FUTURE WORKS

We found that there was no significant change in workload between the two conditions and both the PFM and No-PFM interfaces were found to be usable based on the systems usability survey. Although our results show that there is a performance gain when using the predictive formal modelling feature at runtime, our analysis does not take into account how the accuracy of the prediction could affect the users' performance or trust in the system. Furthermore, this study did not investigate how the PFM feature increases explainability and hence trust in the human swarm interaction. A follow-up study could collect data on trust, acceptability, and user preferences in order to evaluate these measures. In future work, we may also consider embedding recommendations for the operators to help in controlling the swarm, i.e., when to add or remove drones. In order to expand on the assessment of mental workload, different data streams about the users' interaction, such as neurophysiological responses (e.g., error potentials), might be useful. This could indicate the operator's cognitive workload and situational awareness as they operate the swarm in a disaster response scenario.

6 CONCLUSION

Building on previous work in human-swarm interaction on deploying predictive formal modelling (PFM) at runtime, we conducted a within-subject user study to determine its impact on performance and mental workload. 60 participants performed the role of a UAV swarm operator delivering parcels to target locations in a simulation environment. Participants added or removed agents as needed to complete the mission within the given time and budget. Findings show that participants using PFM were able to complete more tasks in less time compared to the No-PFM scenario. This increase did not increase mental workload, showing the benefit of PFM in a human-swarm interaction.

ACKNOWLEDGMENTS

REFERENCES

[1] 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. https://doi.org/10.1109/RO-MAN57019.2023.10309454

[2] Ayodeji Opeyemi Abioye, Stephen D. Prior, Peter Saddington, and Sarvapali D. Ramchurn. 2022. The performance and cognitive workload analysis of a multimodal speech and visual gesture (mSVG) UAV control interface. *Robotics and Autonomous Systems* 147 (2022), 103915. https://doi.org/10.1016/j.robot.2021.103915

[3] 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.

[4] 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.

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

[6] 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.

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

[8] 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. https://doi.org/10.1109/TDSC.2017.2650901

[9] 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.

[10] 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.

[11] 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.

[12] 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. https://doi.org/10.1109/RO-MAN57019.2023.10309626

[13] Sandra G Hart. 1986. NASA task load index (TLX). (1986).

[14] Maria-Theresa Oanh Hoang, Niels van Berkel, Mikael B. Skov, and Timothy R. Merritt. 2023. Challenges and Requirements in Multi-Drone Interfaces. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI EA '23)*. Association for Computing Machinery, New York, NY, USA, Article 65, 9 pages. https://doi.org/10.1145/3544549.3585673

[15] 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.

[16] Panagiotis Kouvaros and Alessio Lomuscio. 2015. Verifying emergent properties of swarms. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*.

[17] Dasom Lee, David J. Hess, and Michiel A. Heldeweg. 2022. Safety and privacy regulations for unmanned aerial vehicles: A multiple comparative analysis. *Technology in Society* 71 (2022), 102079. https://doi.org/10.1016/j.techsoc.2022.102079

[18] 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.

[19] Mingyang Lyu, Yibo Zhao, Chao Huang, and Hailong Huang. 2023. Unmanned Aerial Vehicles for Search and Rescue: A Survey. *Remote Sensing* 15, 13 (2023). https://doi.org/10.3390/rs15133266

[20] 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).

[21] 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.

[22] Prolific. 2023. Prolific - quickly find research participants you can trust. https://www.prolific.co/

[23] 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.

[24] 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. https://doi.org/10.1145/3488242

[25] Mohammad D Soorati, Mohammad Naiseh, William Hunt, Katie Parnell, Jediah Clark, and Sarvapali D Ramchurn. 2023. Enabling Trustworthiness in Human-swarm Systems Through a Digital Twin. *Putting AI in the Critical Loop* (2023).

[26] Usability.gov Improving the User Experience. 2023. System Usability Scale (SUS). https://www.usability.gov/how-to-and-tools/methods/system-usability-scale.html

[27] James Wilson, Greg Chance, Peter Winter, Suet Lee, Emma Milner, Dhaminda Abeywickrama, Shane Windsor, John Downer, Kerstin Eder, Jonathan Ives, and Sabine Hauert. 2023. Trustworthy Swarms. In *Proceedings of the First International Symposium on Trustworthy Autonomous Systems* (Edinburgh, United Kingdom) *(TAS '23)*. Association for Computing Machinery, New York, NY, USA, Article 10, 11 pages. https://doi.org/10.1145/3597512.3599705