Detecting and Classifying Degradation in Robotic Swarms: An Experimental Study

Seth Bullock[1], Jan Noyes[1], Victoria Steane[2],
Chris Bennett[1], Wenwen Gao[1], Sophie Hart[1], Elliott Hogg[1], Debora Zanatto[1]

[1] Faculty of Engineering, University of Bristol, UK
[2] Thales Technology, Research, and Innovation, Reading RG2 6GF, UK
seth.bullock@bristol.ac.uk

Abstract

This paper describes the results of an experiment in which human participants were required to detect degraded robot swarm behaviour and classify it as arising from either faulty or malicious robot activity in an idealised simulation of a multi-agent search and rescue task. The accuracy of participant judgements was influenced by the nature of the degradation, and between-participant differences in the extent to which they interacted with the swarm did not significantly influence their accuracy. It was found that detecting and classifying swarm degradation are challenging tasks that are likely to be strongly sensitive to task setting and will tend to require careful swarm system design and specific operator training.

Introduction

The potential for employing robot swarms to carry out real-world tasks is increasing as technical challenges are overcome and the number of applications grows (Kolling, Walker, Chakraborty, Sycara, and Lewis, 2016). Swarms have many advantages. In a search and rescue application, for instance, they can quickly explore a hostile environment and can continue to function well when individual agents are lost or added to the swarm. For the foreseeable future even a fully autonomous robot swarm will need to be overseen by one or more human operators, and should the swarm's behaviour should become significantly degraded, they may have to take control. This could lead to an Out-Of-The-Loop situation (Endsley, 2017) where although a human operator has a poor understanding of the immediate situation, they nevertheless must take control and recover the swarm. To date, robot swarm research has tended to focus on technological challenges (see, e.g., Vrontis, Christofi, Pereira, Tarba, Makrides & Trichina, 2022) and the role of human swarm operators is still poorly understood (Hart, Steane, Bullock & Noyes, 2022).

One feature of a swarm is that it will degrade. This may be due to robot malfunction, perhaps arising from sensor failure, motor failure or the impact of an undiscovered bug, or may be the result of malicious robot activity, perhaps resulting from hacking of some kind. Thus, it is important that the human operator can respond appropriately to degradation when it occurs.

A fundamental challenge is therefore to determine what factors influence the extent to which operators can detect degradation in swarms, and classify it correctly, for instance as an example of malfunction or malice. Experimental work investigating these questions is reported here.

The Experiment

A simulation of an autonomous robot swarm exploring the interior of an idealised building in two-dimensional continuous space (Hogg, Hauert, Harvey & Richards, 2022; Hogg, Harvey, Hauert & Richards, 2022) was developed and used as the basis for a human-robot interaction experiment. Experimental participants were told that they should help a group of 20 robots to search a building as quickly as possible as a chemical leak had occurred, and incapacitated people may still be in the area (see Fig 1). Participants attempted this task multiple times, completing 15 trials in randomised order: four trials with healthy swarms (i.e., no faulty or malicious robot behaviour was present); three trials with faulty sensor behaviour exhibited by either two, four or six agents; three trials with faulty motor behaviour exhibited by either two, four or six agents; three trials with malicious blocking behaviour exhibited by either two, four or six agents (in which robots blocked doorways to prevent other robots from completing their task effectively); and two trials with malicious communication behaviour exhibited by either one or three agents (in which robots sent deliberately incorrect information to other robots to prevent them from effectively completing the task). In addition to watching the trial unfold, at any time throughout each trial participants could broadcast directional commands to the entire robot swarm ("north", "south", "east" or "west") which all robots would attempt to obey for a short period.

At the end of each trial, participants were asked two questions: Q1 "Based on your experience in the last trial, is it more likely that A) All the robots were working properly, or that B) Some robots were not working properly?"; Q2 "If, in fact, the robots were NOT all working properly, is it more likely that A) Some of the robots were Faulty, or B) Some of the robots were Malicious" to which participants responded using a five-point Likert-type scale ranging from 'A much more likely' to 'B much more likely' with 'A & B equally likely' in the centre.

[1,2] Study Conceptualization and Design: SB, JN, VS, WG, SH, EH, DZ; Coding: EH, SB, CB; Data Analysis: WG, EH, JN, SB; Writing: JN, SB

Before the experimental trials commenced there were three practice sessions. In the first and third, participants observed a healthy swarm completing the task. In the second, participants were instructed on how to control the swarm by pressing arrow keys to command the swarm to move north, south, east or west. Each trial lasted 30 seconds. Fifty-seven participants completed the trials satisfactorily.

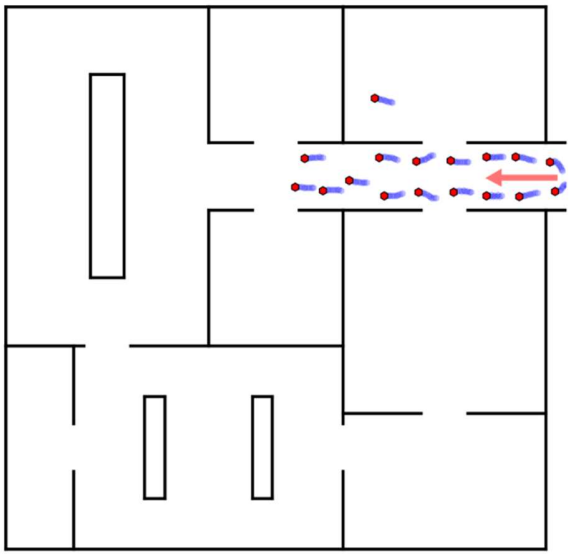

Figure 1. Top-down visualization of a building floorplan and a swarm of autonomous robots beginning to explore it, as experienced by participants during an experimental trial.

Results and Discussion

For trials in which all robots were 'healthy' (i.e., none were faulty or malicious), participants tended to judge that it was either likely or very likely that this was the case (63% of trials) rather than judging it likely or very likely that some robots were not behaving properly (25% of trials) or judging that the two possibilities were equally likely (12% of trials).

Participants had more difficulty identifying cases in which swarms contained *faulty robots*. In 60% of the trials featuring robots with *sensor faults,* participants mistakenly judged that it was likely or very likely that the swarm was behaving properly, only making correct judgements for 27% of this type of trial (and being unsure for the remaining 13% of the trials). For trials containing robots with *motor faults,* participants made correct judgements 54% of the time, but were incorrect or unsure in 35% and 11% of cases, respectively. Moreover, participants were often unable to classify these faulty swarms as suffering from faulty rather than malicious robot behaviour, correctly judging it to be likely or very likely that faults were to blame in 46% of these trials, but judging that it was likely or very likely that malice was to blame in 23% of these trials, and judging both possibilities to be equally likely in almost a third (31%) of the trials.

For trials in which swarms contained *malicious robots* that were attempting to prevent the swarm from carrying out its task, participants were in general more able to detect that something was wrong with 71% of judgements on the correct side of the Likert-type scale versus 18% on the incorrect side. However, classifying the problem was still challenging, particularly for trials featuring *malicious communicators* which were only judged to be likely/very likely to be due to malice in 19% of these trials. By contrast, two-thirds (67%) of trials featuring *malicious blockers* were classified as likely or very likely to involve malicious behaviour.

Overall, when pooled together, distributions of participant responses to healthy trials, faulty trials, and malicious trials differed significantly from each other and from uniform responses for Q1 and Q2 (χ^2, df=4, p<0.0001 in all cases).

The extent to which participant interaction with the swarm influenced the accuracy of their judgements was explored by analysing the relationship between the number of commands given to the swam during a trial and the accuracy of a participant's post-trial judgements. No significant relationship between these two variables was found. Any positive impact on participant judgements stemming from them interacting with the swarm may be being offset by "neglect benevolence" (Walker et al., 2012), i.e., rather than improving their ability to understand what a swarm is doing, interacting with the swarm may also be interfering with an operator's ability to evaluate or classify a swarm's behaviour.

Conclusion

In summary, the results presented here demonstrate that relatively naïve operators of a simulated robot swarm, while able to identify healthy swarm behaviour more often than not, find detecting the presence of different kinds and degrees of swarm degradation challenging. Behaviour resulting from sensor faults was particularly hard to detect and behaviour resulting from malicious communication was particularly hard to classify. By contrast, malicious blocking behaviour was both detected and classified with high accuracy.

There is very little past research on human detection of swarm degradation with which to compare these results. One of the few studies, by Capiola, Hamdan, Fox, Lyons, Sycara and Lewis (2022), asked participants to view and rate 21 simulations of flocking swarms which were subject to different levels of degradation (defined as featuring agents that deviated from the swarm's overall trajectory). They reported that participants struggled to detect degradation, citing the complexity of the task as a possible reason since 256 flocking agents are not easy to monitor.

The results presented here are in line with this previous study in that participants were often inaccurate in their judgements of swarm degradation, even for swarms of only 20 robots. However, the current work goes further in beginning to explore and differentiate between different categories of degradation. Overall, we find that identifying and classifying swarm degradation are both challenging tasks for relatively naïve operators. Our results indicate that the issue of swarm degradation is likely to be strongly sensitive to task setting and, consequently, dealing with it will tend to require careful swarm system design and bespoke operator training.

Acknowledgements. This work was funded and delivered in partnership between the Thales Group and the University of Bristol, and with the support of the UK Engineering and Physical Sciences Research Council Grant Award EP/R004757/1 entitled 'Thales-Bristol Partnership in Hybrid Autonomous Systems Engineering (T-B PHASE)'.

References

Capiola, A., Hamdan, I., Fox, E. L., Lyons, J. B., Sycara, K. & Lewis, M. (2022). "Is something amiss?" investigating individuals' competence in estimating swarm degradation. *Theoretical Issues in Ergonomics Science*, 23(5), 1-26.

Endsley, M. R. (2017). From here to autonomy: Lessons learned from human-automation research. *Human Factors*, 59(1), 5-27.

Hart, S., Steane, V., Bullock, S., & Noyes, J. (2022). Understanding human decision-making when controlling UAVs in a search and rescue application. In *Proceedings of the 7th International Conference on Human Interaction and Emerging Technologies (IHIET 2022)*. Springer.

Hogg, E., Harvey D., Hauert, S., Richards, A. (2020) Evolving robust supervisors for robot swarms in uncertain complex environments. In F. Matsuno, S. Azuma, & M. Yamamoto (eds.) *Distributed Autonomous Robotic Systems: 15th International Symposium*, pp. 120-133. Springer.

Hogg, E., Hauert, S., Harvey D., Richards, A. (2020). Evolving behaviour trees for supervisory control of robot swarms, *Artificial Life and Robotics*, 25, 569-577.

Kolling, A., Walker, P., Chakraborty, N., Sycara, K., & Lewis, M. (2016). Human interaction with robot swarms: A survey, *IEEE Transactions on Human-Machine Systems*, 46(1), 9-26.

Vrontis, D., Christofi, M., Pereira, V., Tarba, S., Makrides, A. & Trichina, E. (2022). Artificial intelligence, robotics, advanced technologies and human resource management: A systematic review. *International Journal of Human Resource Management*, 33(6), 1237-1266.

Walker, P., Nunnally, S., Lewis, M., Kolling, A., Chakraborty, N. & Sycara, K. (2012). Neglect benevolence in human control of swarms in the presence of latency, *IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pp. 3009-3014, doi: 10.1109/ICSMC.2012.6378253.