Hunt, E. R., & Hauert, S. (2020). A checklist for safe robot swarms. *Nature Machine Intelligence*. https://doi.org/10.1038/s42256-020-0213-2

A checklist for safe robot swarms

Edmund Hunt and Sabine Hauert*
Engineering Mathematics, Bristol Robotics Laboratory, University of Bristol

*corresponding author: sabine.hauert@bristol.ac.uk

Standfirst: As robot swarms move from the laboratory to real world applications, a routine checklist of questions could help ensure their safe operation.

Robot swarms promise to tackle problems ranging from food production and natural disaster response, to logistics and space exploration[1–4]. As swarms are deployed outside the laboratory in real world applications, we have a unique opportunity to engineer them to be safe from the get-go. Safe for the public, safe for the environment, and indeed, safe for themselves. This will help build public confidence in their use, and counter hyped or negative narratives about swarms in media and science fiction. Designing safe swarms is also challenging, as the main benefits of swarms, namely their scalability, robustness, and emergent properties, arise from self-organisation, a concept rarely used in engineering[5].

Previous research has identified certain challenges for the deployment of safe robot swarms, particularly in the areas of swarm agent fault tolerance[6-9], human-swarm interaction[8] and swarm security[11–15], but limited consideration has been given to systematic assessment of swarm safety. As a starting point, we propose a preliminary “safe swarm checklist” with 10 questions that should be answered satisfactorily by engineers before a swarm can be deployed in the real world, where real costs are at stake. Highlighting potential risks early in the swarm design phase will allow mitigations to be introduced.

Safety in engineering can be defined as the absence of catastrophic consequences on the user(s) and the environment. It is closely related to concepts of dependability, or the ability to deliver a service that can justifiably be trusted[16]. We take a holistic view of safety that goes beyond analysing failure modes and performing risk analysis[17], to also include the broader socio-technical context of deployment. In our proposed “safe swarm checklist”, questions 1 and 2 on ethics and legality come first as a vital prerequisite for initial testing. Ethical governance and training should be pervasive from the design to the deployment of robot swarms[18]. Questions 3 and 4 relate to accountability and user-swarm interactions. Then, because a defining feature of swarms is their emergent capabilities, we consider individual and swarm-level risks separately for each of the dimensions of physical harm, behavioural harm, and security in questions 5 to 10.

We briefly apply our checklist to a hypothetical swarm of 100 small floating robots – let’s call them bubblebots – deployed to monitor water pollutants (Figure 1). The idea builds on several examples of real-world robot swarm deployments in aquatic environments[19–21]. The bubblebots are meant to distribute over an enclosed floating harbour and light up in ways that communicate their local sensor readings. By sharing information within the swarm, the robots can collectively communicate the overall state of the water in the harbour and reorganise to highlight pollution sources.

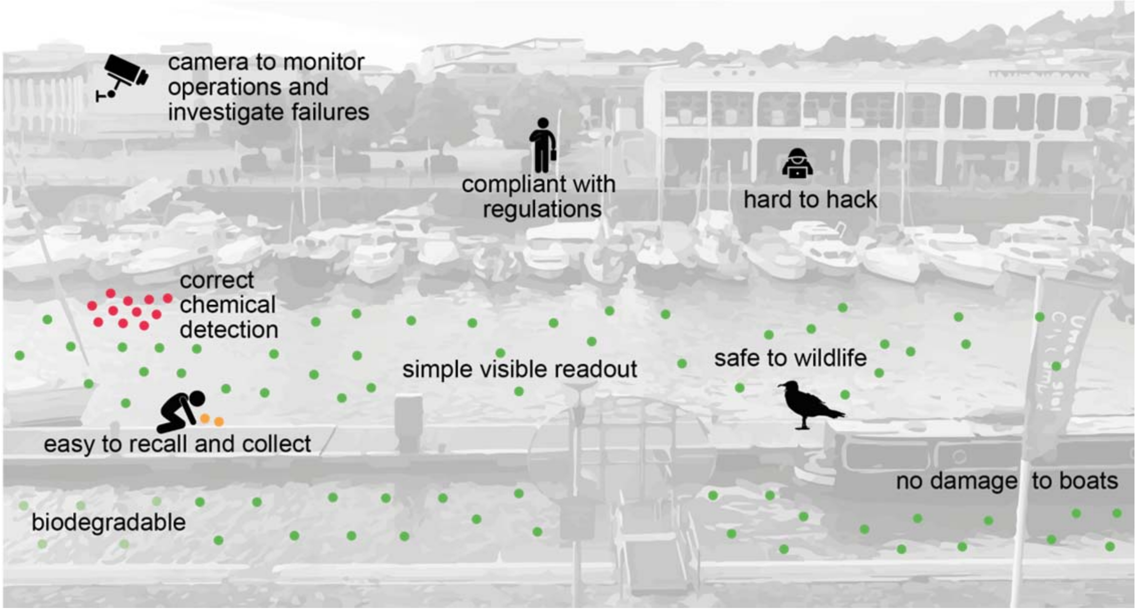

Figure 1 Safety considerations for the deployment of bubblebots used in a floating harbour to signal pollutants.

(1) Ethics. Is this an ethical use of a robot swarm?

We will consult authorities such as the IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems and its work on Ethically Aligned Design[22], or the BSI standard for Ethical Design and Application of Robots and Robotic Systems[23]. With bubblebots, we focus on an application for social good, namely environmental monitoring of water pollutants, considering privacy and potential harm to actors in the harbour. Mutual shaping of the technology between researchers and users will help embed local ethical norms[24].

(2) Legal. Does the swarm comply with all relevant laws and regulations for the domain(s) of deployment?

The bubblebot swarm will need to comply with all relevant rules: environmental, harbour and maritime, or relating to health and safety. There may be a need for public liability insurance.

(3) Accountability. Is there a way to analyse swarm failures?

Following work by Winfield et al.[25], it would be helpful to store short-term recordings of the actions of the robots based on sensory readings and communication in a so-called “black box”, inspired from flight recorders in the aviation industry. This could be done on board the robots, or using an external camera system monitoring overall operations. This information would help to investigate and reconstruct conditions that led to unsafe operations, and would be used to improve swarm implementation if things go wrong.

(4) User interaction. Can the users interact with the swarm to prevent unwanted behaviour?

It should be possible to easily deploy, interact with and retrieve the swarm. In this case, user interaction will involve depositing the bubblebots in the harbour, and reading out the state of the swarm from the harbourside by looking at the robot location and colour status. Bubblebots can easily be stopped using a broadcasted message transmitted throughout the swarm from an operator on the harbourside, in which case robots will home to one area of the harbour for collection.

(5) Physical harm from individual robots. Can the individual robots cause physical harm to humans, animals, or the environment?

Bubblebots will be designed to be small enough to avoid damage to boats in the harbour, or other robots, but large enough to avoid seabirds and fish from eating them. Trials will be done to check that they are compatible with actors in the harbour. They will be buoyant to avoid them sinking and becoming a pollutant themselves, and will be easy to detect for collection by harbour staff. Materials for the waterproof shell will be optimised for durability to avoid breaches, and electronics will be low enough power to avoid possibility of electric shock. Bubblebots failing (power loss, broken sensor or motors) will turn off, avoiding further impact. In the future, bubblebots could even be biodegradable as an additional safeguard – such research is moving beyond the conceptual stage[26].

(6) Physical harm from the swarm. Can the emergent swarm behaviour cause physical harm to humans, animals, or the environment?

The swarm of 100 bubblebots could disrupt natural animal behaviour in the harbour by being a source of distraction, changing their usual feeding habits. Studies will need to be done to assess the impact of the swarm on wildlife. Likewise, the swarm could cause damage to boats or the harbour if they all accumulate in the same location. Algorithmic safeguards will be put in place to avoid dense robot aggregation.

(7) Behavioural harm from individual robots. Can the behaviour of individual robots result in unsafe operation?

Poor programming or lack of consideration of noise in the environment (boats passing by, local disturbance of the sensor from wildlife) may lead individual robot behaviours to display erroneous or unreliable LED colours (constantly fluctuating, or inconsistent with neighbouring robots), which may result in these individual robots unnecessarily worrying the harbourside community and eroding trust in the overall operation of the swarm. Individual behaviours will be thoroughly tested to determine the parameters that lead to stable and reliable signal outputs, and, where possible, the programme will be formally verified to avoid undiscovered use cases[25]. Individual failures can also be detected and signalled by other members of the swarm as a way to make them more visible [6-9].

(8) Behavioural harm from the swarm. Can failure of the emergent swarm behaviour cause unsafe operation?

Faulty swarm operation, either due to faulty individual robots impacting emergent swarm behaviour, or due to poor engineering of emergent properties, may result in incorrect water pollutant assessment. Consequently, pollution could go undetected or false alarms could lead to disruption of harbour operations. Mitigations could include an initial focus on detecting non-safety-critical pollutants that can be easily verified by a human on the ground. For safety critical pollutants, swarm behaviours will either need to be formally verified[27], or tested thoroughly in simulation and reality to gain confidence in the system. A rigorous approval process could take inspiration from the approach used by other sectors, such as the FDA approval process for medicine.

(9) Security of individual robots. Can individual robots be maliciously hacked?

The minimal design of bubblebots will limit the ways in which they can be hacked, including hijacking communication channels, reprogramming the robot controller, or faulting the sensory readings. Securing these potential weaknesses will be a priority. A minimal design will also contribute to privacy, as relatively less information will need to be stored and/or processed onboard each robot.

(10) Security of the swarm. Can the emergent swarm behaviour be subverted by malicious actors?

Swarm behaviours could be subverted by injecting robots with faulty sensory readings into the swarm, or changing the environment, for example by inserting “fake pollutants”. A swarm signature will be added to all bubblebots to ensure they are able to detect internal, versus external actors. Additionally, swarms will aim to communicate unusual patterns in pollutants by displaying a collective “confidence” status using their colour (e.g. orange for unusual activities). One will also need to check whether swarm behaviour can reveal private information, for example through chemical detection near boats, or imaging of personal identifiers.

While this is not meant to be an exhaustive assessment, it provides initial insight into the safety of the swarm. Redundancy in the questions asked, for example behavioural harm leading to physical harm due to poor testing of the harbour water, is intentional and allows for a thorough coverage of safety considerations from different perspectives.

The checklist will identify different risks for different use case scenarios and swarm technologies. Consider applying the checklist to a swarm of robots designed to store and retrieve goods in a warehouse. Swarms can be used ethically in logistics, though amongst broad considerations we will assess their impact on human labour. The swarm will need to comply with regulations in place regarding workplace safety. The user interaction part of the checklist will consider workers in nearby proximity of the swarm unloading or requesting items, those passing by on the shop floor, and supervisors monitoring and controlling the swarm. Such a supervisory system could also allow for short-term recording of the warehouse state, to be used as a black box for accountability if anything goes wrong, or individual robots could locally store a log file for analysis. In relation to physical harm, robots working in densely populated environments with workers, goods, and other robots will need to avoid collisions. Hardware should be designed to be robust to failure, for example detecting sensor or motor malfunction, or battery faults which could cause damage or fires. Collectively, we will need to demonstrate the swarm is able to perform its task without causing physical harm, for example transporting items, without toppling over. To assess behavioural harm, we will consider whether individual robots thoroughly map all possible sensory readings to appropriate actions (e.g. avoiding dangerous full speed motion for example), we will also study the behaviour of the swarm to ensure they don't cause unsafe configurations in the warehouse by blocking exit routes. Security in this scenario might relate to industrial espionage, whereby competitors wish to gain business intelligence about what products are being handled; robots could work effectively without needing to identify the contents of their load. Hackers may also aim to disrupt operations, which would necessitate safeguards to avoid external actors from interacting with the swarm.

Using our checklist, we have begun systematic, albeit theoretical, exploration of safe robot swarm designs for real-world deployment. Designing such swarms is most likely feasible with today's technology and making them thoroughly safe will improve public perceptions in the crucial early trust building stage.

Safe swarms can take many forms, depending on the capabilities of the robots and numbers used. Robots such as bubblebots rely on their simplicity, making them less likely to individually fail in complex ways; less liable to subtle manipulation; and more likely to biodegrade quickly and harmlessly. More capable warehouse robots may instead rely on classical cybersecurity tools and reasoning to make them individually safe. In both cases, swarms should benefit from the philosophy of 'complexity engineering', where we rely on emergence of collective capabilities to get the task done. This puts the focus on getting interactions right, whether within the swarm, with other robot systems or human users, and with the physical world.

The potential for robot swarms to improve our world is enormous: first though, we must build in safety from the beginning. Safe swarms are successful swarms.

Competing interests

The authors have no competing interests to declare.

References and notes

1. Yang, G.-Z. et al. The grand challenges of science robotics. Sci. Robot. 3, (2018).
2. Brambilla, M., Ferrante, E., Birattari, M. & Dorigo, M. Swarm robotics: A review from the swarm engineering perspective. Swarm Intell. 7, 1–41 (2013).
3. Schranz, M., Umlauft, M., Sende, M. & Elmenreich, W. Swarm Robotic Behaviors and Current Applications. Front. Robot. AI 7, 36 (2020).
4. Hamann, H. Swarm Robotics: A Formal Approach. (Springer International Publishing, 2018). doi:10.1007/978-3-319-74528-2
5. Winfield, A. F. T., Harper, C. J. & Nembrini, J. Towards Dependable Swarms and a New Discipline of Swarm Engineering. 126–142 (2005). doi:10.1007/978-3-540-30552-1_11
6. Bjerknes, J. D. & Winfield, A. F. T. On Fault Tolerance and Scalability of Swarm Robotic Systems. in Distributed Autonomous Robotic Systems: The 10th International

Symposium (eds. Martinoli, A. et al.) 431–444 (Springer, 2013). doi:10.1007/978-3-642-32723-0_31

7. Winfield, A. F. T. & Nembrini, J. Safety in Numbers: Fault Tolerance in Robot Swarms. Int. J. Model. Identif. Control 1, 30–37 (2006).

8. Christensen, A. L., O'Grady, R. & Dorigo, M. From fireflies to fault-tolerant swarms of robots. IEEE Trans. Evol. Comput. 13, 754–766 (2009).

9. Tarapore, D., Christensen, A. L. & Timmis, J. Generic, scalable and decentralized fault detection for robot swarms. PLoS One 12, 1–29 (2017).

10. Kolling, A., Walker, P., Chakraborty, N., Sycara, K. & Lewis, M. Human Interaction With Robot Swarms: A Survey. IEEE Trans. Human-Machine Syst. 46, 9–26 (2016).

11. Gil, S., Kumar, S., Mazumder, M., Katabi, D. & Rus, D. Guaranteeing spoof-resilient multi-robot networks. Auton. Robots 41, 1383–1400 (2017).

12. Primiero, G., Tuci, E., Tagliabue, J. & Ferrante, E. Swarm Attack: A Self-organized Model to Recover from Malicious Communication Manipulation in a Swarm of Simple Simulated Agents. in Swarm Intell. (eds. Dorigo, M. et al.) 213–224 (Springer International Publishing, 2018).

13. Higgins, F., Tomlinson, A. & Martin, K. M. Survey on Security Challenges for Swarm Robotics. in 2009 Fifth International Conference on Autonomic and Autonomous Systems 307–312 (2009). doi:10.1109/ICAS.2009.62

14. Sargeant, I. & Tomlinson, A. Maliciously manipulating a robotic swarm. Proc. ESCS 16, (2016).

15. Strobel, V., Castelló Ferrer, E. & Dorigo, M. Blockchain Technology Secures Robot Swarms: A Comparison of Consensus Protocols and Their Resilience to Byzantine Robots. Front. Robot. AI 7, (2020).

16. Avižienis, A., Laprie, J. C., Randell, B. & Landwehr, C. Basic concepts and taxonomy of dependable and secure computing. IEEE Trans. Dependable Secur. Comput. 1, 11–33 (2004).

17. Modarres, M., Kaminskiy, M. P. & Krivtsov, V. Reliability engineering and risk analysis: a practical guide. (CRC press, 2016).

18. Winfield, A. F. T. & Jirotka, M. Ethical governance is essential to building trust in robotics and artificial intelligence systems. Philos. Trans. R. Soc. A Math. Phys. Eng. Sci. 376, (2018).

19. Schmickl, T. et al. CoCoRo -- The Self-Aware Underwater Swarm. in 2011 Fifth IEEE Conference on Self-Adaptive and Self-Organizing Systems Workshops 120–126 (2011). doi:10.1109/SASOW.2011.11

20. Thenius, R. et al. subCULTron - Cultural Development as a Tool in Underwater Robotics Consortium for coordination of research activities concerning the Venice lagoon system. Artif. Life Intell. Agents Symp. (2016).

21. Duarte, M. et al. Evolution of Collective Behaviors for a Real Swarm of Aquatic Surface Robots. PLoS One 11, e0151834 (2016).

22. The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. Ethically Aligned Design: A Vision for Prioritizing Human Wellbeing with Artificial Intelligence and Autonomous Systems. (IEEE).

23. British Standards Institute. BS 8611:2016, Robots and Robotic Devices: Guide to the Ethical Design and Application of Robots and Robotic Systems. (2016).

24. Carrillo-Zapata, D. et al. Mutual Shaping in Swarm Robotics: User Studies in Fire and Rescue, Storage Organization, and Bridge Inspection. Front. Robot. AI 7, (2020).

25. Winfield, A. F. T. & Jirotka, M. The Case for an Ethical Black Box. in Towards Autonomous Robotic Systems (eds. Gao, Y., Fallah, S., Jin, Y. & Lekakou, C.) 262–273 (Springer International Publishing, 2017).

26. Rossiter, J., Winfield, J. & Ieropoulos, I. Here today, gone tomorrow: biodegradable soft robots. Electroact. Polym. Actuators Devices 2016 9798, 97981S (2016).

27. Dixon, C., Winfield, A. F. T., Fisher, M. & Zeng, C. Towards temporal verification of swarm robotic systems. Rob. Auton. Syst. 60, 1429–1441 (2012).