# Safety and Desirability Constraints for Machine Learning Systems – Workshop Report

David Bossens[1], Marina Konstantatou[2], Saina Akhond[2], Sergio Araujo-Estrada[1], and Shane Windsor[3]

[1]University of Southampton
[2]Foster+Partners
[3]University of Bristol

Report date: October 25, 2022

Workshop date: October 19, 2022
Workshop venue: Foster+Partners, Riverside, 22 Hester Rd, London SW11 4AN

## 1 Introduction

The workshop on Safety and Desirability Constraints for Machine Learning Systems takes place within the context of the Integrator project *Safety and Desirability Constraints for AI-controlled Drones on Construction Sites* which takes place in the context of the Trustworthy Autonomous Systems network (see `https://www.tas.ac.uk/`), a multi-institute research programme focusing on issues around trustworthy autonomous systems. The project joins together University of Southampton as part of the TAS Hub and the University of Bristol as part of the Functionality Node of the TAS Network. The project considers applications of AI-controlled drones for the transport and monitoring applications on construction sites, focusing particularly on implementing desirability constraints. Desirability constraints would help to avoid financial, societal, ecological, or bodily harm, protect privacy and security, and ensure transparency to ensure the drone's behaviour displays intention and capability. Such technologies come with stakeholders from a variety of sectors including construction, security and defence, insurance, and logistics. The project will investigate flight simulators in complex construction environments, implement desirability constraints into control, design suitable sets of models for robustness to uncertainty, perform lab-based tests in miniature urban environments, as well study how to interact with human operators through alerts and interventions. The project's industrial partner, Foster+Partners, which is the UK's largest architectural firm, supports the project in various ways including hosting this very workshop.

The main purpose of the workshop was to inform the project on the societal relevance and the design of the constraints. The participants first introduced themselves to each other and then the Integrator project was briefly explained in a 15 minute presentation. Then followed two group discussions, including a general session about interests and application areas in constrained machine learning and a scenario based session which considers specific applications of drones in construction sites.

The workshop included 20 participants who actively participated in group discussions, including the organisers (David Bossens, Marina Konstantatou, Saina Akhond, Sergio Araujo-Estrada, and Shane Windsor), Liam Braeger (Katapult), Suet Lee (University of Bristol), Sina Sareh (Royal College of Art), Mohammad Divband Soorati (University of Southampton), Josef Musil (Foster+Partners), Mohammad Naiseh (University of Southampton), Laetitia Regnault (Expleo), Rasoul Sadeghian (Royal College of Art), Govind Muthukrishnan (Expleo), Shahrooz Shahin (Royal College of Art), Tao Sun (Foster+Partners), Georgios Athanasopoulos (Foster+Partners), Charlotte Jones (Katapult), David Silverstone (Liverpool Victoria), and Simon Withers (University of Greenwich).

## 2 General session

Participants were interested in a wide variety of applications, including human-robot ineraction; construction worker safety; interpreting big data; autonomous nuclear inspection and space based infrastructure monitoring; clinical decision-making, human-in-the-loop and explainability; 3D printing via robot swarms, ML-based simulation and image generation; control of robotic grippers and identification of people; bio-inspired flight dynamics and control of uncrewed air vehicles; manufacturing; digital twins; obstacle avoidance; urban scaling; digital heritage; supervised predictive modelling for car insurance; IOT devices to detect defects in infrastructure; agriculture; mining; ML-based design of morphologies; safety and resilience in robotics; natural language processing. The participants mentioned the

plethora of risks in these applications, including uncertainty in the environment (e.g. unanticipated events, cluttered environments, lack of feedback), privacy and security issues (e.g. unwanted data sharing, cyberattacks and hacking), corrupted data, real-time performance as well as bias and fairness. The AI system could be constrained to avoid collisions with obstacles or humans, take into the limited memory on-board, respect rules, norms, and regulations, to limit autonomy and enforce gradual exploration, to constrain the environment to be more predictable, to keep the agent in sight of the human operator, and to shape the morphology of the system to better be suited to the local environment and control algorithm. Strategies to enforce these constraints may include prediction and monitoring (e.g. digital twin, performance metrics), using soft materials or other hardware configurations, clearly defining safe areas and pathways for the AI and clear norms and regulations for humans, using information from multiple modalities, conventional computer vision and identification mechanisms, human oversight (diagnosis, surveys, monitoring, documentation, and solution engineering), sharing information between multiple agents, switching between levels of autonomy, and making the task easier and simplifying the environment.

# 3 Scenario based session on drones in construction sites

In the scenario based session, participants of the workshop were asked to answer questions about a particular scenario related to monitoring or transport applications of drones in construction sites. The questions concerned an analysis of stakeholders; identifying AI capabilities;machine learning techniques and data sets; safety concerns as well as ethical, legal, and societal challenges; the constraints to address these concerns and challenges, how to enforce these; and the types of interventions when something does go wrong.

## 3.1 Moving construction materials on-site

**The scenario** Consider the FB3 from FlyingBasket depicted in Figure 1, which is cargo drone that can take up to 100kg as a payload. The FB3 is promising construction workers to move construction materials in a faster, more flexible, and efficient way while saving on the cost of cranes. The FB3 is traditionally controlled by a human operator. However, in this scenario, we will consider how to control the drone automatically based on machine learning techniques (learning could be online and/or offline).


Figure 1: The FB3 cargo drone from FlyingBasket (see `https://flyingbasket.com/`).

**Responses** Stakeholders of this scenario include the construction site operators, site management, and hod carriers. This scenario also came with a specific question about what kind of materials would be unsuitable for transport, the response included liquids, long objects, and objects subject to high air friction. Important AI capabilities included situational awareness, good communication and collaboration between drones, sampling of environmental contexts/properties that is representative of the real world, accurately modelling the environment, and sensing and perception capabilities. To implement these capabilities, data sets or simulators should accurately model wind, air pressure, payloads, centre of mass of objects, etc. in a suitable physics simulator as well as take into account safety concerns and human feedback (human-in-the-loop). The need for continual learning and transfer learning is also highlighted. In terms of constraints, participants mentioned a) LOLER and PUWER regulations should be satisfied (which can be formulated by a chartered engineer); b) extensive testing of all components before use, particularly for risk of breaking and the risks associated with the components being attached to the drone; c) formulating a lone-working policy to make sure it's effective to work alone. Interventions include alerting for dangerous events (strong wind, human presence, collisions, low power); pre-defined mitigation strategies; and human intervention.

## 3.2 Monitoring building progress

**The scenario** Automated building progress reports may be done by periodically (e.g. on weekly or monthly basis) recording informative videos across the building site, scanning for the progress of different components (e.g. complete,

near completion, just started, yet to be started) and verifying the accuracy of components that are supposed to be complete or near completion. This is especially true when the construction site is very large and the project lasts for extended periods of time, as is the case in the warehouse construction example mentioned in Figure 2. In this scenario, we will look at automating this process as much as possible with machine learning, including the control of the vehicle as well as when and what to monitor, especially when the warehouse construction takes place in more crowded environments.


Figure 2: Left: news headline on construction monitoring (see `https://thedronelifenj.com/projects/`). Right: future projects may also consider warehouse in more crowded environments.

**Responses** The group mentioned a wide variety of stakeholders, including architects, the construction company, subcontractors, regulators, neighbours, general public, offsite manufacturers and material suppliers. Benefits of such technologies include improved error detection, progress updates, reduced inspection costs, and improved information sharing. Costs include safety risks, privacy, and noise. AI capabilities needed include measuring geometry, understanding sensory data, detecting anomalies, interpreting plans, path planning, obstacle detection, and automated design. As machine learning methods, one may consider supervised learning based on sensory data and CAD models, reinforcement learning based on simulation and lab data, active learning, and imitation learning to learn from experts (e.g. humans or conventional computer algorithms). Concerns about safety included fall hazards, while desirability criteria included predictable drone behavuiours, communication with workers, privacy of neighbours and workers, and security concerns. Mitigating these concerns may be done by constraining the time of operation, using privacy filters on collected data, and clearly communicating hazards and other issues with workers.

## 3.3 Power plant safety monitoring

**The scenario** Installing new structures onto an existing power plant can involve serious safety risks and there is a need for continually monitoring such risks including when the crew are working on site. The headline shown in Figure 3 shows how real world projects are currently using drones for recording crew operations and inspection of structural components. In this scenario, we will look at automating this process as much as possible with machine learning, including the control of the vehicle as well as when and what to monitor.


Figure 3: News headline on power plant safety monitoring (see `https://thedronelifenj.com/projects/`).

**Responses** A variety of stakeholders were mentioned for power plant safety monitoring, including working crew, energy plant owners, insurers, government agencies or inspectors, suppliers, solution providers, and neighbours. Benefits include improved productivity, safer working conditions, accessibility (as those tasks may be too dangerous for humans). Costs include maintenance of drones, privacy measures needed, unexpected failures of the drone, additional staff to ensure the technical execution is safe and maintains privacy. In terms of AI capabilities, one may consider using sensors to record various safety-relevant data (e.g. images, gas emissions, temperature, ultrasound, laser, sound) to monitor the safety status across the site and connect these to the Internet of Things as well as integrate these sources with data integration. This may be combined with anomaly detection to record data selectively upon anomalous events. Other important capabilities are mapping the building and detecting who are the workers. Reinforcement learning or other machine learning techniques would be good to have more flexible decision-making.

For anomaly detection and classification, Bayesian techniques may be able to detect outliers by providing probability distributions of events at particular times of the day and deep neural networks such as convolutional networks and LSTM networks would be beneficial to analyse video data. A potential pipeline may involve to collect data about safe and unsafe power plant sensory readings and then performing batch training on these data to detect unsafe states.

### 3.4 Monitoring tresspassers

**The scenario** On construction sites there is often a need for protecting against unauthorised access. In this scenario, we will look at drones that automatically position, monitor, and alert against unauthorised access using machine learning techniques. To help imagine the scenario, consider the BeeHive system (see Figure 4), which uses machine learning and other techniques for the following capabilities:

> *A lightweight drone built for autonomous flight, the Bee features high-quality video capture, rapid recharging, and reliable landing even in adverse weather conditions. It is able to respond to activity on any part of the property, within 30 seconds on a 4-acre property and 90 seconds on a 10-acre property. The Bee operates by autonomously planning a safe flight path around the pre-mapped property, detecting unexpected obstacles, and performing a "security sweep" before landing back in the Hive, all without any need for manual drone operation.* — DroneLife

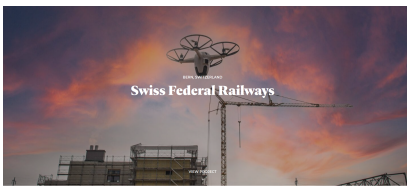


Figure 4: The BeeHive from Sunflower Labs (see `https://sunflower-labs.com/`).

**Responses** Beyond traditional stakeholders associated with the building sector and monitoring applications, the security staff, workers, and neighbouring individuals are most affected by the tresspasser monitoring scenario. Important AI capabilities include distinguishing trespassers from workers, identifying the intent of the tresspasser (e.g. vandal, unintentional), coverage of the area without missing spots, mapping and localisation, detection-of-motion and identification, obstacle avoidance, intent detection, sentiment analysis from movement, and deterrence. Machine learning methods to solve these include dynamic SLAM methods for localisation and mapping, neural network for detection and identification, as well as path planning or reinforcement learning for high-level control. Safety concerns include the drone startling the workers or the trespasser, integrity of the construction site, as well as interference in high-risk operations around the construction site. Desirability criteria include biases in identification, privacy of neighbours, and alerts that are not too annoying. Staying within the construction site is a hard constraint or soft constraint (depending on if there is a barrier), workers should be identified correctly to not raise an alarm, and worker annoyance is a soft constraint. In terms of interventions, users can alert the drone to return to safest way out and override misidentification. Distraction strategies, in which one trespasser distracts the drone on one location, may require multiple drones scanning a part of the environment and communicateing each other's findings.

## 4 Conclusions

The workshop brought together insights from industry and academia to discuss exciting machine learning applications, consider the wider context of society in the formulation of AI technologies, as well as think of constraints, interventions, costs, and risks in using drones on construction sites. Apart from these immediate benefits, the workshop also helped establish new communication channels among people with common interests, and the organisers hope that that all participants had a great time and that the event brought about long-lasting memories.