Safety assurance of autonomy to support the Fourth Industrial Revolution

A briefing paper commissioned by Policy Links (Institute for Manufacturing, University of Cambridge) on behalf of the Global Manufacturing & Industrialisation Summit (GMIS) and Lloyd's Register Foundation (LRF)

Richard Hawkins and John McDermid
Assuring Autonomy International Programme
University of York

July 2019

Lloyd's Register Foundation

GMIS Global Manufacturing & Industrialisation Summit

About the Briefing Paper

This is one of two briefing papers developed under the project "The safety and security dimensions of Industry 4.0", commissioned to the University of Cambridge (Policy Links – Institute for Manufacturing) by the Global Manufacturing & Industrialisation Summit (GMIS), on behalf of the Lloyd's Register Foundation (LRF).

The briefing papers constitute the first stage within an open, multi-stakeholder project that aims to bridge the safety and security knowledge gaps in the deployment of 4th Industrial Revolution (4IR) technologies in manufacturing. The intention of the briefing papers is not to be prescriptive, but to provide an in-depth analysis of selected themes related to emerging safety and security risks and requirements for manufacturing, in the context of 4IR.

This briefing papers has been produced without formal GMIS or LRF editing. The views expressed here do not imply the expression of any opinion on the part of GMIS or LRF. Mention of firm names or commercial products does not constitute an endorsement by the authors, GMIS or LRF.

ASSURING AUTONOMY INTERNATIONAL PROGRAMME

The Assuring Autonomy International Programme is advancing the safety of robotics and autonomous systems (RAS) across the globe. It is a £12million partnership between Lloyd's Register Foundation and the University of York that is working with an international community of developers, regulators, researchers and others to ensure the public can benefit from the safe, assured and regulated introduction and adoption of RAS. The Programme is addressing core technical issues underlying the assurance of RAS, supporting industrial demonstrator projects, delivering training and education, and creating an online Body of Knowledge that will reflect the evolving state-of-practice in assuring and regulating RAS.

GMIS | Global Manufacturing & Industrialisation Summit

The Global Manufacturing and Industrialisation Summit (GMIS) was established in 2015 as an industry association to build bridges between manufacturers, governments & NGOs, technologists, and investors in harnessing the Fourth Industrial Revolution's transformation of manufacturing to the regeneration of the global economy. A joint initiative by the United Arab Emirates and the United Nations Industrial Development Organization (UNIDO), GMIS is a platform that presents the manufacturing sector with an opportunity to contribute towards global good, working to the benefit of all.

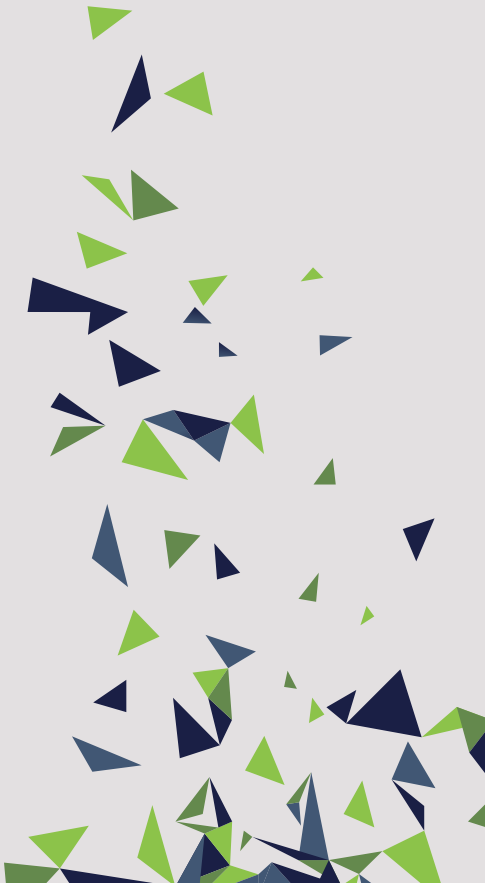
Lloyd's Register Foundation

The Lloyd's Register Foundation is a UK charity established in 2012. With our mission to protect the safety of life and property, and to advance transport and engineering education and research, the Foundation has an important role to play in meeting the challenges of today and the future. Our vision is to be known worldwide as a leading supporter of engineering-related research, training and education that makes a real difference in improving the safety of the critical infrastructure on which modern society relies. In support of this, we promote scientific excellence and act as a catalyst working with others to achieve maximum impact.

Key messages

The Fourth Industrial Revolution will lead to a step change in the level of autonomy in manufacturing facilities. Autonomy involves enabling machines to make decisions that were previously made by humans. This movement of responsibility for decision-making from humans to machines raises a number of challenges when assuring the safety of the operations. The challenges are exacerbated further by the use of artificial intelligence in system training and control, and by the greater use of robots in areas where there is close interaction between humans and machines (often referred to as cobots).

In this paper, we first provide a general discussion of the challenges of autonomy, and then we use three scenarios for manufacturing, representing a progression through different levels of automation and autonomy. This highlights how the nature and extent of the safety assurance challenge increase for each scenario. We discuss possible approaches for addressing the issues and identify the key open research challenges.



1. Introduction to the importance to the Fourth Industrial Revolution of safety assurance of autonomy

The advent of the Fourth Industrial Revolution (4IR) has many potential benefits in terms of efficiency and cost-effectiveness of factory operations. However, it also brings challenges relating to our ability to assure the safety of these more advanced manufacturing facilities. Where the operation of a system may adversely impact safety, it is necessary to show there is sufficient assurance that the system is safe to operate prior to deploying that system into operation. By assurance, we mean justified confidence in the properties and behaviour of the system. While assuring the safety of factory operations is always challenging, 4IR brings additional challenges and considerations that must be addressed. Many of these arise from the introduction of increased autonomy – the movement of responsibility for decision-making from humans to machines. The challenges of 4IR are exacerbated where artificial intelligence (AI) is used in system training and control, and by the greater use of robots in areas where there is close interaction between humans and machines (often referred to as cobots).

At the most fundamental level, assuring the safety of systems requires us to:

1. Provide a clear and unambiguous definition of how the system must behave in all situations that the system might encounter during operation in order to be considered safe.
2. Implement the system such that it provides the required behaviour and generate evidence to demonstrate this.
3. Gain a detailed understanding of things that might go wrong when the system is operating, identify if these might affect safety, and demonstrate that sufficient mitigation has been put in place for those things.

In the following we refer to these as aspects of assurance. Each of these aspects is challenging in its own right. There are accepted approaches to safety assurance that have developed over several years, supported by regulations and standards, in order to increase clarity on what is expected and to simplify compliance. For example, in Europe one of the key instruments is the Machinery Directive,[1] which places requirements on "machines" – an assembly fitted with a drive system where at least one part of the system moves. The Directive invokes standard safety principles, such as identifying hazards and risks. There are also similar requirements in other parts of the world, for example, from the Occupational Health and Safety Administration (OSHA) in the USA. Often there is a national-level regulator, for example, the OSHA in the USA and the Health and Safety Executive (HSE) in the UK, that enforces the rules.

Standards bodies, such as the International Organisation for Standardisation (ISO), produce safety standards, for example, ISO 13857,[2] which sets "safety distances" for protecting people from hazards. Generally, these standards impose requirements for simple physical barriers (see the scenarios in Section 3 below for more details).

1 Directive 2006/42/EC of the European Parliament and of the Council of 17 May 2006 on machinery, and amending Directive 95/16/EC.

2 1SO 13857 First Edition 2008-03-01, Safety of Machinery – safety distances to prevent hazard zones being reached by upper and lower limbs.

Factories have long seen increases in automation, to reduce manning and to increase throughput, quality and repeatability (among other things). The concept of computer numerical control (CNC) machines dates back to the 1940s, with CNC milling machines being widely used in some manufacturing processes. Simple factory robots (often no more than an arm with some specific tools attached) have become widespread, for example, for welding and painting car bodies. Generally, the safety of such systems is still achieved and assured through the use of physical separation of people and machines – such as guard fencing.

Factory machines are now becoming more sophisticated, and there is an increasing desire to introduce so-called cobots, which work closely with human operators. This can be seen as yet another evolutionary step in factory automation – but it is one that brings increasing challenges – and we must consider why such systems cannot be addressed within existing safety assurance practice and regulatory frameworks.

First, many of these systems, for example, cobots, now have humans and machines working in close physical proximity, where physical separation is not a viable option – indeed, in many cases, the robot and the human need to physically touch in order to carry out their joint task. Thus, more subtle forms of safety mechanisms, for example, force limitation, are needed.

Second, as systems become more autonomous – meaning that what would otherwise be human decision-making is transferred to a machine – it is much more difficult to predict and assess the safety of the machine. This is in contrast to automated systems where the behaviour, for example, the trajectory of a cutting head, is known and predictable.

In this paper, we first provide a general discussion of the challenges of autonomy, and then we use three scenarios for manufacturing, representing a progression through different levels of automation and autonomy. This highlights how the nature and extent of the safety assurance challenge change for each scenario and discusses possible approaches for addressing those challenges. We also identify the key open challenges and related needs in research and in education and training.

2. The challenges of autonomy

The introduction of autonomy into a factory setting involves enabling machines to make decisions that were previously made by humans. Decisions made by humans in a factory are often very complex, relying on the operator's understanding of the outcomes of decisions that are made and how those outcomes may be influenced by the current situation in the factory. Many decisions are often made based on the judgement, experience and instinct of the operator, which ensure that the decisions result in a safe outcome. The safety assurance for such decisions arises from the skill, training and contextual awareness of trusted operatives. Decisions that are made by a machine whose consequences may impact safety will also require assurance, which must be based upon explicit evidence about the behaviour of the machine in the different situations it might encounter during operation.

This is a challenge even when the environment is constrained and static. It becomes even harder when the autonomous system must operate within an open, or unconstrained, environment. In addition, the more interaction the autonomous machine has with other things in the environment, the harder safety assurance becomes to achieve. The machine might be required to interact with other machines, or to interact with humans. All machines operating within a factory setting will need to interact with humans in some way. Even robots that are deemed to be fully autonomous will normally have to share their space with humans, even if only during exceptional periods such as for maintenance.

One approach for assuring autonomous systems is therefore to try to constrain the environment in order to make it easier to provide guarantees about the machine's safety. In factories this is a very common strategy that has been adopted – constraining the environment by installing physical barriers. This provides a controlled and predictable working environment for the machine, but it also crucially prevents humans and other machines from interacting with the machine in uncontrolled ways. This approach, while making safety assurance much easier to demonstrate, and essentially enabling traditional approaches to safety assurance to be adopted, also limits the amount of autonomy that can be utilised. To fully realise the benefits of autonomy in factories, such constraints on the operating environment of the machine need to be relaxed to allow flexibility, and to enable humans and machines to work together. This, in turn, makes the safety assurance challenge harder and requires new approaches and techniques.

Many of the tasks that are being automated in factories, particularly in open and dynamic environments, can be complex in nature. This can make it very difficult to describe precisely how the machine must behave in every possible situation it might encounter. Therefore, an approach that is increasingly being considered is instead to get the machine to learn how to behave through training using a technique known as machine learning (ML). Machines in factories could be trained either by allowing them to operate in the factory setting under close supervision, or by using a simulation. Normally a combination of real-world and simulation is used to provide training data. Once the machine has been trained well enough, it should then be able to cope safely with the situations it encounters in the factory. There are a number of important assurance challenges concerning the use of ML, which we discuss in more detail later.

3. Scenarios for autonomy and safety assurance in 4IR

To explain the issues introduced in Section 2 more fully, we discuss three automation scenarios. The first two scenarios are relatively simple, representing the progressive automation of factories. The third scenario is more challenging, reflecting the issues arising from fully autonomous systems and cobots. However, as almost all modern systems are computer controlled, we emphasise some of the technical issues and what this means for assurance.

In the discussion we focus mainly on the third aspect of safety assurance introduced above. In particular, we illustrate hazards – situations that, if uncontrolled, might lead to harm – and the controls for the hazards.

3.1 Scenario 1 — physical barriers, fixed operation

Scenario description

In this scenario we consider two distinct examples of fixed operations: milling machines and production line robots. The term “fixed operations” does not mean that the machines cannot be repurposed, but rather that their purpose is likely to be defined stably for a period of time, with explicit intervention needed to change their function. However, in the examples used here the machines are fixed in the sense of remaining in the same physical location.

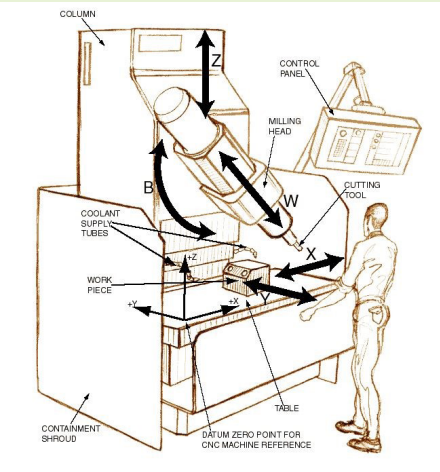


Figure 1 - CNC milling machine

Most modern milling machines can produce complex-shaped components by removing material from a “blank” – perhaps a cuboid metal block. Such machines have interchangeable cutting heads and can turn both the work piece (the object being cut) and the cutting head. For many years factories have used computer numerical control (CNC) milling machines, where the cutting path and actions are defined by a computer program. Figure 1 shows a CNC milling machine,[3] indicating the axes of movement – but with an operator to give an idea of scale. Modern

3 Illustration courtesy of www.efunda.com

machines do not have operators in attendance, and they usually have more degrees of freedom in terms of moving the workpiece, as well as the tools.

Safety assurance considerations

Focusing on the safety perspective, the hazards generally relate to the production of high-energy debris – parts of the work piece or tools "flying off" and striking people in the vicinity. Protection "systems" will generally be simple. Should the hazard occur, there will be physical containment around the machine, as can be seen in Figure 1. Hazard prevention includes keeping operators away from the machine and powering off the machine prior to undertaking maintenance activities. More subtle prevention mechanisms might include validation of the program to show that there are no unintended interactions between the tool and work piece (this is assurance aspect 2) – for example, lateral force on a drill – and perhaps force feedback sensors leading to the machine being powered down, if forces near safe limits.

Production line robots (see Figure 2) can be used for a range of tasks, for example, assembly, welding, painting; generally, the robots can be programmed to do different tasks, and the same type of robot may be used at multiple points on a production line, each programmed for a particular task. These robots are often powerful, capable of moving fast and with high forces. From a safety perspective, the hazards generally involve physical contact between the robots and people – potentially causing temporary or permanent injury. As well as physical impact, injury could also arise from the effects of hot tools, for example, welding heads. Typically, the hazard is controlled by preventing humans being in the area when the robots are operating, for example, through fencing with limited access points, and ensuring that the robots are powered off if the access doors are opened (see Figure 3).

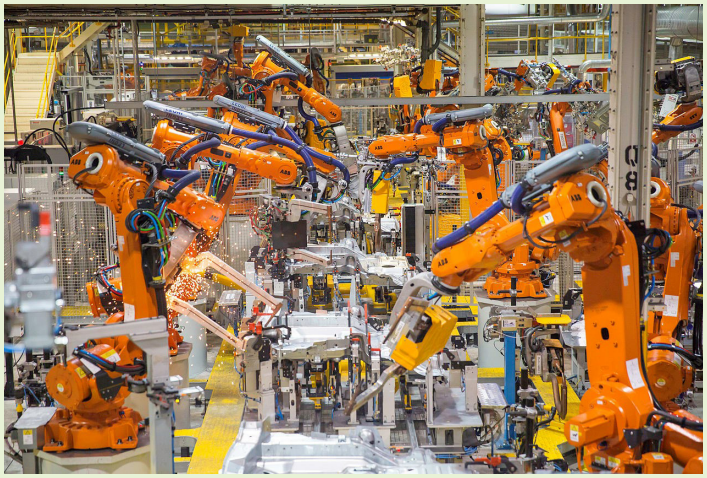


Figure 2 - Production-line robots

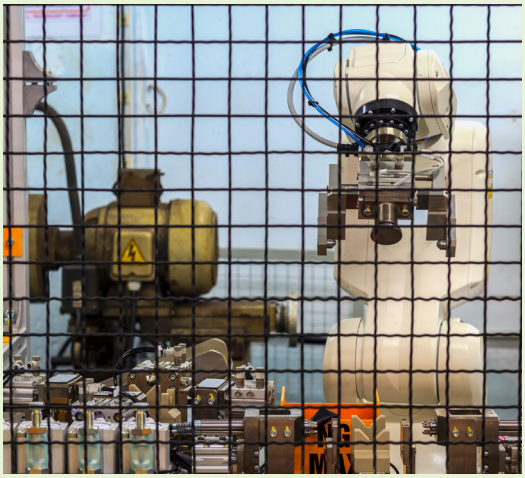


Figure 3 - Robot with safety fencing

Additionally, hazards might arise from high-energy debris, and this will be controlled by ensuring that fences are high enough to prevent debris from going over the top (or enclosing the machines above as well), by ensuring that the fence will contain debris up to an assessed energy level, and so on. Some of the standards may help in defining fencing characteristics.[4]

3.2 Scenario 2 – virtual barriers, flexible operation

Scenario description

In this scenario, we consider two aspects of flexible operation: first, the machines move (are not at a fixed location), and second, the purpose may change. For brevity, we focus more on the movement of the machines.

We consider, as an example, automated guided vehicles (AGVs) that move blank work pieces and partly machined parts around a factory. In essence the AGVs move from work station to work station, carrying the parts. Thus, they might take blanks from goods inwards or a warehouse to a first machine, move partially finished products between machines, and move completed parts to an assembly shop or a dispatch centre.

A typical AGV will be an electric vehicle, with sensors, for example, barcode readers, to help it navigate and, say, infrared sensors to detect obstacles. A typical AGV is shown in Figure 4. An overview of a simple factory operation using AGVs is shown in Figures 5 and 6. There will be mechanisms for moving the work piece between the AGV and the workstations. AGVs might have mechanisms to achieve/support the transfer, or this may be done entirely by mechanisms at each workstation.

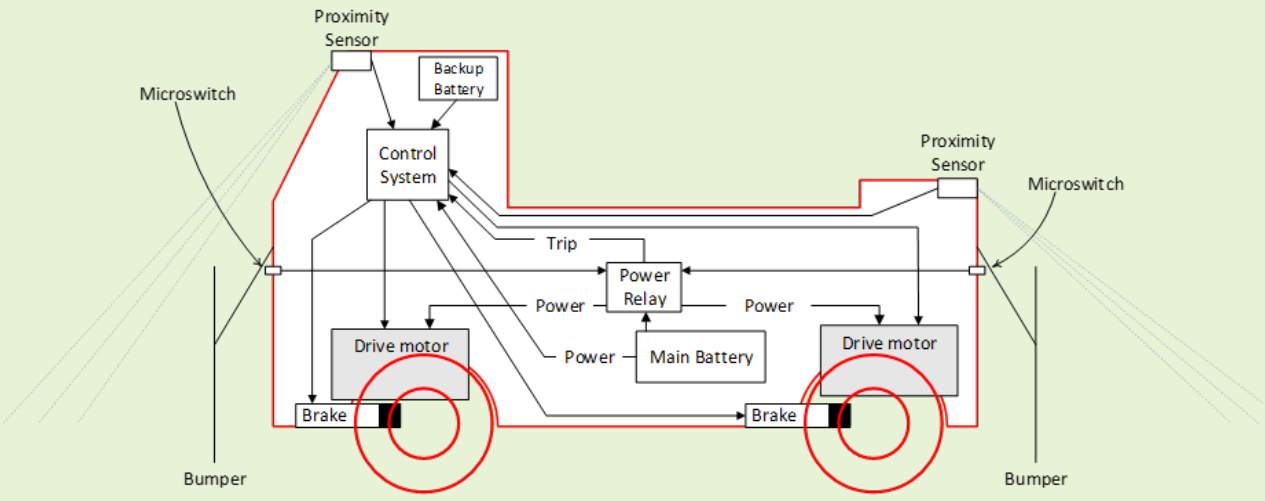


Figure 4 - Simple AGV with electric motor, sensors and computer control system

4 EN ISO 14120 Safety of Machinery guards. General requirements for the design, construction and selection of fixed and movable guards.

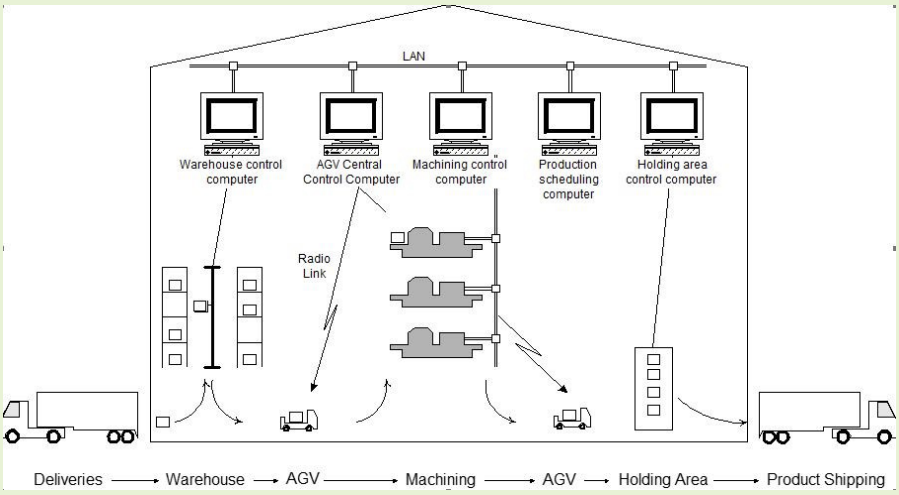


Figure 5 - AGVs used to move materials around a factory

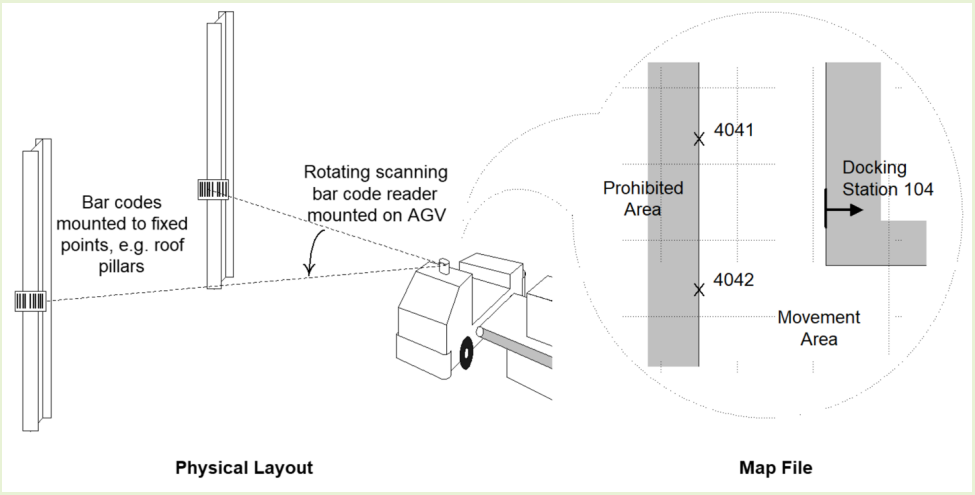


Figure 6 - AGVs navigating using digital maps and barcode readers for localisation

Safety assurance considerations

The main hazards relate to impact with people causing injury. This can happen through hitting people moving through the factory. However, probably the worst case is if someone is impacted by a falling workpiece or the AGV itself (an AGV may weigh half a tonne or more). This might occur if an AGV falls over an unprotected edge, for example, a loading bay, and lands on people below. More minor hazards might relate to burns from hot parts being carried by the AGV, or possibly spills of fluids such as battery acid. There are also problems relating to damage to other machines or equipment through impact.

As a result of the nature of the AGV operation, hazard control via physical barriers, for example, fences,

is not practicable. There are several possible control mechanisms for the main (impact) hazards. A key avoidance mechanism is verification of the planned paths through the factory to ensure that they avoid static objects (noting that maps must be updated if the factory layout changes). In terms of mitigation, if a hazard does arise the AGV can be fitted with sensors to detect impending impact and apply the brakes, and soft-impact bumpers that might both apply the brakes and power off the motors.

There will also be value in using mechanisms that are not onboard the AGV to assist in hazard control. For example, light curtains (see Figure 7) can be used to warn of proximity to an unprotected drop in time to apply the brakes. More subtly, geofencing can be used to mark off "exclusion zones" on digital maps that the navigation software would avoid in moving around the factory. Note that this starts to emphasise the first and second aspects of safety assurance – how we specify clearly the "prohibited areas" in the digital map, how we verify that the software for operating the AGV determines location and assesses whether or not the trajectory will breach the prohibited areas and, if so, how the emergency stop (applying brakes and powering off motors) is triggered. While, for modern systems, this is not complex it does involve a substantial amount of software development, analysis and testing. A range of standards are relevant,[5,6] which require extensive verification and validation of the software, addressing assurance aspect 2. A significant change between scenario 1 and scenario 2 is the increased emphasis on, and cost of, the software development and evidence-generation activities.

If the purpose of the factory is changed, then the roles of the individual machines may alter – and there may also be a change in the set of machines employed. If the AGV system has been designed well, the hazard-control mechanisms will work equally well with the new machines and activities, and little extra development and verification work will be needed. Perhaps the biggest issue will be checking (validating) new digital maps to ensure that the "safe navigation" functions are appropriate for the new layout. Note that safety of data is a key issue, and there is now emerging guidance on such issues.[7]

Figure 7 - Light curtain

As a final observation, safety processes must consider the whole operation. As AGVs will be designed to stop if there is a problem, consideration needs to be given to how they will be restarted or "rescued". Is this to be done manually? If so, how, and what are the hazards of sending workers onto the factory floor? If it is to be done remotely, how do the remote operations interface with the automated functions of the AGV, can remote operations over-ride the emergency braking, and so on? All of this is standard safety engineering good practice – and should be well within the competence of an experienced professional.

5 IEC, 2010. 61508 ed2.0 – Functional Safety of Electrical/Electronic/Programmable Electronic Safety-Related Systems. International Electrotechnical Commission (IEC).

6 RTCA, 2012. DO178C – Software Considerations in Airborne Systems and Equipment Certification. Radio and Technical Commission for Aeronautics (RTCA).

7 The Data Safety Initiative Working Group (DSIWG), Data Safety Guidance version 3.0, SCSC-127C, 2018, available at: https://scsc.uk/r127C:2

3.3 Scenario 3 – Cooperative working, use of ML

Scenario description

Increasingly, robots are being used for tasks where they need to cooperate closely with humans; hence, the physical separation used in the previous two examples cannot be achieved if the robots are to be effective. This shift is coupled with the use of ML for training the robots, and both trends bring with them challenges for safety and assurance. For simplicity, the scenario is split into two, considering training robots offline, that is, prior to operation, and more dynamic operation where they adapt their operation to the behaviour of an operator.

One method of training robots is for a human operator to "demonstrate" the performance of tasks to a robot, which is then able to learn how to perform the tasks itself. Where the tasks are fixed and performed in exactly the same manner each time, training the robot is straightforward, requiring simple mimicking and repetition. For example, a robot manipulator that might be trained to perform tasks is shown in Figure 8.[8]

Figure 8 - Robot manipulator

In situations where the robot must perform tasks in a much more dynamic environment, such as where humans are working in proximity to the robot, it is not possible to demonstrate to the robot how to perform every task in all situations (the number of situations the robot may encounter are too large and too uncertain). The robot must instead be capable of learning from specific examples how to perform the tasks in a much more general manner. ML is a technology that can enable robots to learn how to perform tasks in dynamic environments involving interactions with humans.

Safety assurance considerations

As before, the hazards arise from physical interaction between the operator and the robot at sufficient speed or force to cause injury. There are two aspects to assurance. First, will the robot reliably follow a safe path? Second, will the robots stop (quickly enough) to prevent harm to the operator if there is an impact?

For the first consideration, assuring a safe path for the robot is largely an issue of the integrity of the software, and the accuracy of what has been learnt. For the simple case, the training process will involve an operator carrying out the designated task or tasks multiple times, with the robot "averaging" from the training sessions to produce the moves that will be implemented in operation.

8 Image courtesy of MCM

For the more complex, dynamic case involving the use of ML the assurance process needs to consider the sufficiency of the training data sets. This is a surprisingly complex issue. If the robot does not see enough training data, it is less likely to learn correctly how to perform the task safely. Conversely, too much training data may lead to "overfitting", where the robot is very good at dealing with situations it has encountered during training but is unable to safely adjust its behaviour to new situations. There are strategies that can help to deal with this, such as computing an "envelope" within which the robot can move (this might be done statistically), but determining what training data is good enough from a safety assurance perspective remains an open challenge. Assurance also needs to consider the underlying software on which the control algorithms rest – this might be a software package for ML, for example, Google's Tensorflow,[9] and will include software for reading sensors, controlling actuators, and so on.

The use of software packages such as Tensorflow or Microsoft Cognitive Toolkit[10] pose problems, as the development process is not (normally) visible to the robot manufacturers or to regulators. Here assurance arguments have to be based on arguments that the software has been "proven in use". This is known to be problematic[11] and standards may impose limitations on what can be claimed. There are also issues of software updates (see Section 3.4 for more discussion). The assurance of control software for sensors and actuators is "conventional" and can be carried out using the standards identified above, for example, IEC 51508.

The second consideration is whether the robots stop (quickly enough) to prevent harm to the operator if there is an impact. It is likely that "emergency stop" functions will be programmed explicitly, and this will be subject to standard software assurance techniques, as discussed above. In practice, the most difficult part might be to determine precisely what should be considered hazardous behaviour, taking into account the range of physiology of the operators, the effects of personal protective equipment (PPE), the size of workpieces that might be used, and whether or not workpieces are sufficiently hot that even momentary touching could cause burns or cuts, and so on. Validation of the specifications is likely to involve simulation, as "direct" validation would put operators at risk.

Where robots and operators engage in more dynamic collaboration the machines are often referred to as cobots – or collaborative robots. The example in Figure 9 below shows cobots on a highly automated, but flexible, production line. The cobots operate completely autonomously during production but interact with operators for retooling (this is not within the competence of the machines). Both the operator and the cobot move in a confined space, and the trajectories of movement can intersect – hazards again relate to contact between the cobot or its end-effector (tool) and the operator. There will need to be an "emergency stop" capability for the cobots – similar to that just described. The assurance challenges come in the functions for avoiding impact.

9 https://www.tensorflow.org/

10 https://docs.microsoft.com/en-us/cognitive-toolkit/

11 Littlewood, B. and Strigini, L., 1995. Validation of ultra-high dependability for software-based systems. In Predictably Dependable Computing Systems (pp. 473–493). Springer, Berlin, Heidelberg.

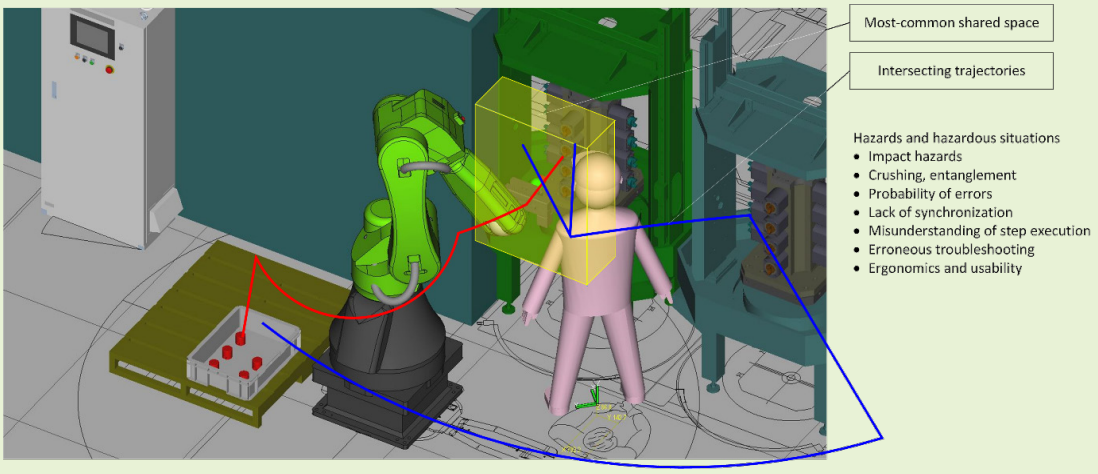


Figure 9 - Cobots in a factory setting

In the example of Figure 9,[12] the cobots carry out trajectory prediction, both for their own movement and for that of the operators. If they predict that the paths will violate a minimum separation distance, then the cobot will re-plan its movement. This is likely to be more subtle than an emergency stop, for example, involving reducing speed or taking a longer path, to ensure that the system is able to operate effectively. Furthermore, the minimum separation is likely to be small, as the operator may need to "hand" a tool to the cobot, meaning that the distance is less than the overall dimensions of the tool.

An underlying challenge here is how to train the cobot in a safe manner. In order for the cobot to learn how to avoid impact with humans, an effective approach may be to use a technique known as reinforcement learning (RL).[13] This is a form of ML where the cobot learns by being "rewarded" for taking safe decisions, and "penalised" for taking unsafe decisions. Note that this approach requires the cobot to make unsafe decisions while training in order to learn to avoid such decisions during operation. It would of course not be acceptable to train the cobot by letting it conflict with operators in the real world. It is therefore likely that simulation will be used to enable exploration of the operational space, and to train the cobot to avoid hazardous situations without exposing operators to additional risk. The assurance then rests on the soundness of the training environment, including the accuracy of the simulation in modelling human, as well as cobot, movement, and the software used to evaluate satisfaction of safety objectives. The success of the approach is also dependent on the correctness of assumptions made in the simulation about operator physiology and behaviour (e.g. speed of movement), and the behaviour of the cobot may not be safe if the operators do not behave in the way that is modelled. It may be possible for the operators to wear non-conventional PPE, that is, fabrics with embedded transponders, so that the cobot can determine the operator's position accurately. This would reduce the problems of invalid assumptions about operator movement (but this is not currently standard technology, and would need to be introduced sensitively to the operators).

An alternative to using simulation to train the cobot is to try to assure safety "dynamically", by adjusting the cobot's path based on predicted conflicts of trajectories. This involves changes in behaviour while the cobot is operating, and this leads us to the concept of "dynamic assurance", where assessment and

12 Image courtesy of MCM.

13 Sutton, R.S. and Barto, A.G., 2018. Reinforcement learning: An introduction. MIT Press.

assurance of the safe behaviour are conducted during operation, namely, online not offline, and this is likely to involve either or both of:

- Run-time verification[14] – checking that the planned trajectories remain safe as they evolve, in a similar way to guiding RL with safety objectives, but done at run-time not offline;
- Dynamic safety or assurance cases[15] – updating the safety case with information from operation, for example, on achieved separation distances between the cobot and the operator and providing alerts if safety margins are being eroded.

Neither of these approaches are standard practice, although both are the subject of ongoing research.

RL is an established technique; however, its use for safety-related applications is the subject of much ongoing research as a result of the number of potential difficulties. A widely cited paper on "concrete problems of AI"[16] describes a number of example problems such as the notion of "reward hacking", which means meeting some overall objectives in learning behaviour while ignoring other aspects of desirable behaviour. With our cobot example, if it learns to move on the shortest path between two points (optimising time and energy usage) but ignoring the need to avoid conflicts in the overlap zone shown in Figure 9, then this is potentially unsafe. As a result of the complexity and opacity of the learning process, it is difficult to have confidence that such "reward hacking" has not occurred, as it may only affect behaviour in some rare circumstances and hence not show up in system testing.

This scenario clearly highlights that what is achievable in factories with the latest cutting-edge technology far exceeds what it is currently possible to assure. The safety and assurance issues that arise with this scenario are not within the competence of typical safety engineers. Furthermore, the online assurance activities blur the boundaries between the traditional responsibilities of safety and design engineers. There are also issues for regulators, as the assurance approaches being developed and deployed are outside their experience. Thus, there is a major education and training need here. Finally, it should be noted that there are major changes in work patterns for operators, again identifying a need for education and training, and their inclusion in the proposed introduction of these technologies.

3.4 Connectivity and cross-cutting safety assurance issues

This paper has illustrated the issues related to the safety assurance of robots and autonomous systems by considering some progressively more complex scenarios, with the later ones dependent on ML. This progressive approach was taken to expose the range of assurance issues, but there are also some cross-cutting concerns that could apply to all the scenarios. Many of these cross-cutting concerns arise from the requirement to connect systems to a computer network. In order to implement the capabilities discussed, it is inevitable that the robots will be network-connected at least some of the time, and in practice most will be connected to the Internet, either directly or indirectly. In particular, many modern factories make use of "Internet-of-Things" (IOT)[17] technology to increase connectivity. This

14 Calinescu, R., Weyns, D., Gerasimou, S., Iftikhar, M.U., Habli, I. and Kelly, T., 2017. Engineering Trustworthy Self-Adaptive Software with Dynamic Assurance Cases. IEEE Transactions on Software Engineering.

15 Denney, E., Pai, G. and Habli, I., 2015, May. Dynamic safety cases for through-life safety assurance. In 2015 IEEE/ACM 37th IEEE International Conference on Software Engineering (Vol. 2, pp. 587–590). IEEE.

16 Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J. and Mané, D., 2016. Concrete problems in AI safety. arXiv preprint arXiv:1606.06565.

17 Telecommunication standardization sector of ITU. Overview of the Internet of things, Y.2060 edition, 6 2012.

raises a number of unique safety assurance challenges.[18] Here we focus on two such issues.

First, there is the issue of security that arises once any system becomes connected to a computer network. Specifically, we are concerned with the impact that the exploitation of security vulnerabilities arising from the connectivity might have on the safety of the robot. Safety and security have similar objectives, being concerned with understanding and controlling the negative impact of system weaknesses, and there are shared concerns relating to software integrity. However, the specific techniques used in the two disciplines are quite different, and it is common for safety and security work to be carried out independently, by two different sets of specialists.

There has been work on "unifying" safety and security over the years, particularly focused on the safety impact of security exploits.[19] In essence, these approaches involve considering the role that security exploits might have in causing hazards, or in undermining hazard controls. While there is some practical experience in using these techniques,[20] their use is not (yet) common practice, and again there is an issue of education and training, as safety specialists do not normally have a good understanding of security, and vice versa. Moreover, there is growing understanding that the safety impact of security concerns increases with autonomous systems, perhaps most obviously due to the limited opportunity for human oversight/control of any security problems. Although there is research in this area it is far from mature.

Second, there is the issue of software updates, including "over the air" updates for systems that are connected via wireless networks. This can be seen most simply with the AGV example – if the factory layout changes, and new maps are sent to the AGVs, how can we be confident that all the AGVs are working with the new maps, and how can we assure synchronisation so that the AGVs do not use the new maps prematurely? In essence, this can be done by having suitable protocols around map updates, and the assurance process would need to focus on the soundness of the implementation of the protocols, including what happens if updates are lost, corrupted, and so on. This is not unfamiliar territory for safety engineers and should not be too challenging if changes to factory layout are infrequent and involve "down time" while machines are moved or reprogrammed, among other things.

However, as updates become more dynamic, they become harder to assess and assure for safety. Robots that update their control software during operation, for example, may share what they have learnt with other robots operating in the same factory (thus improving the performance of the entire fleet of robots). The underlying issue here is one of impact analysis – how "big" is the change, and how far does its influence extend? Many updates might be benign – having a positive impact on safety, or at least not having a negative impact – but how do we determine this? The more dynamic (frequent) the updates, the harder this is to do.

It should also be considered that if robots learn during operation then they might all change their software in different ways based on what they have learnt. As well as being challenging for safety assurance, this is a problem for regulation, which normally approves all systems of a particular type, rather than each system independently. This is acknowledged as a problem in some domains, for example, automotive, but there are no widely recognised solutions to this problem in the public domain.

18 Jaradat, O., Sljivo, I., Habli, I. and Hawkins, R., 2017, September. Challenges of safety assurance for industry 4.0. In 2017 13th European Dependable Computing Conference (EDCC) (pp. 103–106). IEEE.

19 S. Kriaa et al., "A survey of approaches combining safety and security for industrial control systems", Reliability Engineering and System Safety, 139 (2015) 156–178.

20 I. Fovinoa, M. Maseraa, A.Cianb, "Integrating cyber attacks within fault trees", Reliability Engineering & System Safety, Volume 94, Issue 9, September 2009, pages 1394–1402.

4. Open challenges and areas for action for the safe adoption of 4IR

The discussion above has identified a number of challenges – they are briefly summarised here, followed by some general issues that should be addressed to meet current and future industrial needs.

- Assurance of ML – gaining confidence in the use of ML to provide critical functions (including decision-making) for factory robots, whether learning is offline or during operation;
- Run-time verification – assuring the safety of behaviour during operation, where it is not practical to do so offline;
- Dynamic safety cases – updating safety cases with information from operations so they accurately reflect the robot behaviour;
- Safety impact of security – providing an effective and practical means of assessing and assuring safety where security weaknesses might contribute to hazards;
- Management of software updates – carrying out impact analysis to ensure that any system changes have a known and acceptable impact on system safety;
- Agile assessment and regulation – responding rapidly to changes in systems and software so that the assurance artefacts and regulatory approvals reflect the current system status.

In some respects the last point is over-arching – most of the other points provide a means of enabling agility in assessment and regulations.

Tackling these challenges will require industry, researchers and regulators to work together in the following key areas.

Application-specific research

Many of these issues require research, particularly research with an application focus. There is significant interest in the verification of ML, for example, formal analysis and testing of neural networks (NN), but much of the emphasis is on image analysis, as this is of interest in many domains such as autonomous driving. This is not particularly relevant in the context of advanced manufacturing, where the focus is more upon decision-making for tasks such as path and motion planning. In addition, much of the research on developing machine-learning algorithms focuses on demonstrating small improvements in performance, which are often less important for safety than the ability to, for example, reduce false-negative classifications. Thus, research is needed that is more focused on the needs of factory automation.

Education and training

There are education and training needs in safety assurance for professionals working in robot development to bring people up to the 'state of the art' in assurance of robotics and to highlight the areas where there remain unknowns. This includes training for:

- Engineers – to ensure that safety assurance is properly considered during system development
- Operators – who need to interact with the systems
- Management and board members – to ensure that appropriate governance takes place
- Regulators and assessors – to ensure effective assessment and oversight of autonomous operations

This will ensure that the risks and uncertainties of the evolving technology are taken into account when designing and assessing systems. Working with operators to understand the information that they need to have available (visual, audible, tangible, etc.) will also ensure that safely-operating cobots are accepted, with appropriate information available to operators to engender justified confidence in their use.

Safety assurance standards

There needs to be work on updating standards to make them more relevant to future factory automation. While there is some work underway, more needs to be done to reflect the capabilities of even the current technology. More fundamentally, standards development needs to become much more "agile", namely, the ability to respond to changes in technology and systems design more quickly. Typically, the development of standards takes years, and some, for example, the first issue of IEC 61508, was in development for more than a decade. If standards cannot be updated more rapidly, they will become irrelevant. This is a challenge for standards bodies, which might find their role challenged, for example, by the introduction of more community-based good-practice guidelines. Greater agility is needed in standards-making, and some standards bodies recognise the challenge and are seeking to work more quickly, for example, by developing Publicly Available Specifications (PAS) to set out initial guidance on a topic as a precursor to developing a full standard.

Approach to regulation

In the same way as for standards, the process of regulation must also become more agile to reflect the pace of change in system development and deployment, and this must include dealing with issues such as over-the-air updates. Regulators also face a shortage of the relevant skills resulting from the rate of change of technology. This is exacerbated by the fact that it is more attractive for those with the relevant knowledge and skills to work in industry developing and deploying, rather than regulating, the technology.

4.1 Key areas of action for the safe adoption of 4IR technologies in manufacturing

With the advancement of robotics, the 4IR, and around the assurance of these systems, especially cobots, the manufacturing sector needs to investigate how to safely adopt these technologies.

Assurance strategies and arguments

A critical challenge for adopting modern robotics, using ML, is how to assure them and to gain acceptance of the systems. A three-phase project, carried out across the sector, could help to address this problem; the project would be best carried out by a working group set up with representatives from a range of manufacturing organisations, covering the spectrum from small-scale specialist developers to the operators of large facilities. The first phase should review the RECOLL project (see Figures 8 and 9) to understand how this particular application of cobots has been assured. Second, the sector should produce sector-specific arguments for assurance about cobots, and manufacturing robotics more generally, building on the AAIP BoK (see Section 6) and template assurance arguments produced by the Assuring Autonomy International Programme. Third, the working group should identify the appropriate forms of evidence for supporting the arguments. If appropriate, this should be documented as an industrial guideline, noting that industry can move much faster than standards bodies.

Industrial best practice

The work on assurance would focus on gaining approval for 4IR products, but it would not address the development of such products. These developments will need extensive tool support, for example, for developing and testing the control software. As with the work on assurance, a working group could seek to define and document best practice for development. Here the working group should include developers and users of the 4IR technologies, but also include tool suppliers, especially those working in other safety-related sectors that are already addressing some of these problems. The aim would be to gain cross-sectoral requirements on development methods and to identify requirements for tools. This would both help the developers and users of cobots and other 4IR products, and also serve to stimulate the tools' suppliers to develop more relevant products, by giving clearer requirements to work to. Considering the inclusion of groups at the interface between academia and industry, such as the Advanced Manufacturing Research Centre, will enable the working group to consider best practice across different equipment manufacturers and developers, providing a wide perspective.

Safety and security

Historically, techniques and methods for assessing safety and security have evolved independently, although, as noted, there is some work on drawing the techniques together, including assessing the impact of security weaknesses and vulnerabilities on system safety. The sector could work together to provide practical and focused guidelines on assessing safety and security. A valuable perspective is how to take safety and security into account early in the development life cycle when it is possible to make trade-offs between safety and security, in the context of the system being developed. This is a little-researched area, but it is important to managing safety and security in a cost-effective manner. Some early work on this topic[21] might provide a useful starting point for a working group. This work might usefully be merged, in time, with both the development guidelines – to help in producing safe and secure systems – and the assurance strategies and arguments, to offer a basis for providing broadly based assurance in systems.

21 Asplund, F., McDermid, J., Oates, R., Roberts, J. (2018) Rapid Integration of CPS Security and Safety, IEEE Embedded Systems Letters https://doi.org/10.1109/LES.2018.2879631

5. Conclusions

The advent of 4IR has many potential benefits in terms of efficiency and cost-effectiveness of factory operations. However, it also brings challenges. Many of these arise from autonomy itself – the movement of responsibility for decision-making from humans to machines. The challenges are exacerbated by the use of artificial intelligence in system training and control, and by the greater use of robots in areas where there is close interaction between humans and machines (often referred to as cobots), and so on.

Addressing these problems is a research issue, but there are wider issues too. The robots being developed are outstripping the skills and experience of safety professionals and regulators so there is a need for focused education and training for this community, and work on standards, as well as assessment methods.

Thus, a broad-based set of initiatives is needed in order for industry to realise the potential benefits from the 4IR. The Assuring Autonomy International Programme (AAIP), funded by the Lloyd's Register Foundation, is addressing some of the key issues as a cross-sectoral concern – with some effort focused on advanced manufacturing. Thus, the programme can assist in solving the problems, but more work is needed that is focused explicitly on the 4IR to enable the benefits to be realised, safely, in the near term. The Programme also has the remit to influence standards and regulations, and to provide education and training, and thus is intended to address many of the challenges identified here.

6. Suggested resources

The AAIP is developing a Body of Knowledge (BoK)[22] to reflect the evolving state-of-practice and state-of-the-art in assuring and regulating robotics and autonomous systems across domains. Thus, in time, the BoK should include material that is directly relevant to advanced manufacturing, for example, based on the cobots illustrated in Figure 9. It will also include material developed in other domains, for example, on the interaction of security and safety, which may be relevant to the sector.

In the UK, the Health and Safety Executive (HSE) publishes extensively on machinery and equipment safety,[23] and the HSE's guidance may also be relevant in other countries, as a result of the (current) harmonisation with European legislation.

The UK's Manufacturing Technologies Association (MTA)[24] is a long-standing trade body in the UK, which provides information on a range of manufacturing technologies. The MTA's Standards Update provides information on the fast-moving set of standards in the sector, including identifying work on safety standards.

Information on the safety of manufacturers' robots can be found on their websites; Kuka[25] addresses issues regarding the safety of cobots directly, and other manufacturers, for example, ABB,[26] also provide safety capabilities for their factory robots. Note that the reference to these two manufacturers is intended to identify relevant information, not to endorse specific products.

22 https://www.york.ac.uk/assuring-autonomy/body-of-knowledge/
23 http://www.hse.gov.uk/work-equipment-machinery/index.htm
24 https://www.mta.org.uk
25 https://www.kuka.com/en-gb/products/robotics-systems/software/hub-technologies/kuka_safeoperation
26 https://new.abb.com/products/robotics/controllers/irc5/irc5-options/safemove-2

About Policy Links

Policy Links is the knowledge exchange unit of the Centre for Science, Technology & Innovation Policy (CSTI), University of Cambridge. It aims to provide professional advice and education services grounded in the latest academic research to address the needs of policy officials and civil servants working in the areas of technology, manufacturing and innovation policy.

Policy Links is part of IfM ECS, a wholly owned subsidiary of the University of Cambridge. IfM ECS is embedded within the Institute for Manufacturing (IfM), a division of the University of Cambridge Engineering Department.

Policy Links | IfM Education & Consultancy Services | University of Cambridge | 17 Charles Babbage Road | Cambridge CB3 0FS

+44(0)1223 766141 | www.ifm.eng.cam.ac.uk/services/policy-links

---

UNIVERSITY OF CAMBRIDGE

IfM Education and Consultancy Services

POLICY LINKS

Safety Assurance of Autonomy to Support the Fourth Industrial Revolution