Global Initiative For Industrial Safety

Cambridge Industrial Innovation Policy

UNIVERSITY OF CAMBRIDGE

Assuring the safe deployment of AI-enabled robotics in industrial contexts

RICHARD HAWKINS AND JOHN MCDERMID, UNIVERSITY OF YORK

About this briefing paper

This briefing paper is part of a series developed under the Global Initiative for Industrial Safety (GIFIS), led by Lloyd's Register Foundation, Cambridge Industrial Innovation Policy (CIIP), and the United Nations Industrial Development Organization (UNIDO).

This paper uses the Balanced, Integrated and Grounded (BIG) framework for AI safety cases, developed by the Centre for Assuring Autonomy at the University of York. Specifically, it draws on this framework to analyse the safety implications of deploying AI-enabled robots in forefront industries.

Drawing on expert input, the paper is intended to inform practitioners, policymakers, researchers, and technology developers interested in the safe deployment of AI-enabled robotics in industry.

This paper has not undergone formal editorial review by Lloyd's Register Foundation or GIFIS, and the views expressed do not necessarily reflect the official positions of these organisations. References to firm names or commercial products do not imply endorsement by the authors, GIFIS, or Lloyd's Register Foundation.

Authors

Richard Hawkins, University of York and John McDermid, University of York.

Contents

Key messages....3

1. Safety Assurance of AI-enabled robotics....4
2. AI Ethics Argument....6

   Examples of AI Ethics Arguments for AI-enabled Robotics....7
3. AI System Safety Argument....12

   Examples of AI System Safety Arguments for AI-enabled Robotics....16
4. Purpose-Specific AI Model Safety Argument....19

   Examples of Purpose-Specific AI Model Safety Arguments for AI-enabled Robotics....21
5. General-Purpose AI Model Safety Argument....24

   Examples of General-Purpose AI Model Safety Arguments for AI-enabled Robotics....26
6. Conclusions....28

Key messages

- **AI-enabled robotics can improve industrial safety, but they also introduce new assurance challenges**
  Robotics can remove workers from hazardous tasks and improve operational flexibility, but AI creates new forms of uncertainty, including bias, opacity, unpredictable behaviour, and difficulties in assigning responsibility.
- **Safety assurance must consider the whole system, not the AI model in isolation**
  The safety of AI-enabled robots depends on their operating context, system architecture, human roles, organisational setting, and interaction with other hardware and software components.
- **The BIG Argument provides a structured framework for AI safety cases**
  The Balanced, Integrated and Grounded (BIG) framework links ethical considerations, system-level safety assessment, and AI model assurance into a coherent safety case.
- **Ethical considerations are central to AI-enabled robotics**
  Safety assurance must address not only physical harm, but also issues such as human autonomy, transparency, accountability, fairness, and the distribution of risks and benefits across workers, operators, and maintenance staff.
- **Different AI-enabled robotics applications require different assurance strategies**
  The factory autonomous vehicle, robot arm, and humanoid robot examples show that assurance approaches must be tailored to the task, operating environment, level of autonomy, and type of AI being used.
- **Special-purpose AI and general-purpose AI raise different safety questions**
  Purpose-specific AI models can often be assured through defined requirements, testing, robustness analysis, and deployment controls. General-purpose AI, including large language models, is harder to assure directly and may require architectural safeguards such as input/output guardrails and controlled operating domains.

1. Safety Assurance of AI-enabled robotics

Recent advances in artificial intelligence (AI) are accelerating a new wave of robotics deployment across various industries. While this represents a significant opportunity, it also introduces a challenge: how can we use these technologies while still being able to assure their safety. Although AI brings unique safety assurance challenges, it is crucial that in trying to address these new issues we don't lose sight of the core safety assurance principles and best practices that have been established over decades, and whose effectiveness has been demonstrated through safe operation of industrial systems.

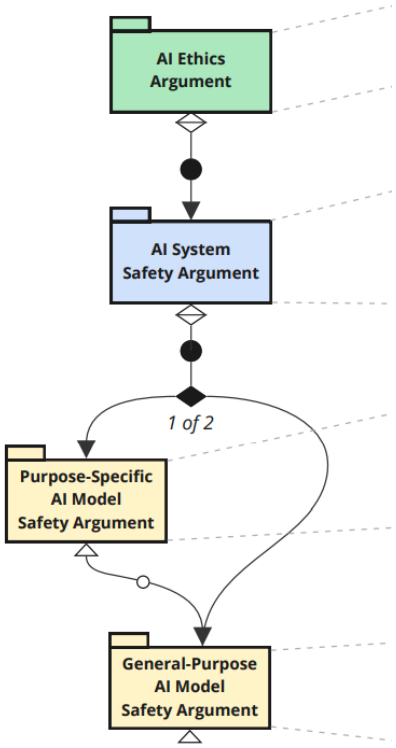
A successful approach to AI safety assurance must be **Grounded** in these existing principles. One of the key principles is that safety is a property of the system as a whole; it can only be understood, assessed, and managed by considering the entire system. Consequently, it is impossible to discuss the safety of AI independently of its system and operational context. At a pragmatic level, a robot arm used for moving items from one conveyor belt to another has a very different set of risks to one used in operation of a sheet metal press. Thus, assuring the safety of AI-enabled robots requires understanding how AI contributes to hazards when performing specific tasks within a particular system in a given operating environment. This necessitates an **AI System Safety Argument** that considers the operational context and the system's ability to operate safely within, and potentially beyond, the defined context. The focus is on claims and assumptions concerning the overall system architecture (such as redundancy, diversity and monitoring), hazard/risk analysis and the definition of safety requirements.

The principle of proportionality is also well-established in safety, aiming to justify the sufficiency of the safety risk posed by the system throughout operation. For AI systems, arguments of sufficiency become more challenging due to the wider ethical considerations that are at play, and the trade-offs that must be considered and accepted by the stakeholders. Therefore, the approach must be **Balanced**, ensuring safety is weighed alongside other concerns, especially ethical issues. This further requires an **AI Ethics Argument** that positions and balances system safety within a wider set of ethical concerns.

Furthermore, the safety of AI models themselves must be **Integrated** with these broader system and ethical considerations. There must be explicit traceability between the safety requirements of the overall system and the specific safety properties of the AI models. From an assurance perspective, it is important to distinguish between what we would refer to as **Purpose-Specific AI Models**, which are specifically trained to undertake specialist tasks as part of a particular system (such as a neural network trained using supervised learning to recognise objects in a defined environmental context), and **General-Purpose AI Models**, such as Large Language Models (LLMs), that have been developed to provide a diverse set of general tasks and capabilities. Even for general-purpose AI models, safety cannot be claimed generally; it can only be demonstrated within a specific operational and system context. For this reason, general-purpose AI models are often adapted for specific system applications; some assurance of the general-purpose model can be done "upstream", but safety can only be evaluated "downstream" in a system and operational context. Ultimately, safety assurance for AI-enabled systems in industrial contexts is achieved by considering a range of safety arguments that together provide a Balanced, Integrated, and Grounded (BIG) safety case.

We illustrate the BIG argument below. The figure shows four argument modules which, together, form an assurance argument for the system in its operational context. In this report we consider each of these argument elements in turn, within an industrial context.

FIGURE 1 SHOWING THE RELATIONSHIPS BETWEEN THE DIFFERENT ARGUMENTS REQUIRED IN AN AS SYSTEM SAFETY CASE.



In overview, the AI ethics argument considers the balance of benefits and harms across stakeholder groups, e.g. showing that the use of AI-enabled robotics doesn't unreasonably re-distribute risks from operators to maintenance staff. This module also helps to set safety requirements on the system as a whole. The system safety argument shows how the system meets the safety requirements "flowed-down" from the level of the ethical arguments, reasoning about both operation in the intended context of use, and how it deals with excursions from the intended context, e.g. due to lighting failing. The two lower-level argument modules then address safety of purpose-specific or general-purpose AI within the system and operational context. In practice, in developing a system and the assurance arguments for it there will be iteration across the levels. In the illustrations here we describe the concepts as if there is a clear "flow-down" from the ethics argument but then consider iterations when we discuss the lower levels.

2. AI Ethics Argument

This report considers the safety of robotic industrial systems that use AI. Safety is not the only property of an industrial system however that must be assured prior to its deployment. Clearly systems must be performant and reliable if they are to be useful in industrial settings. This inevitably leads to trade-offs which must be made and justified. The tension between safety and performance is well known for industrial systems, where measures that can increase safety can also have a negative impact on performance. Consider as an example the widespread use of interlocks and safety cages; whilst these approaches are highly effective at reducing the risk exposure of workers, they can also delay operations and maintenance where worker access is required. Where such tensions exist, a judgement is required on the acceptability of the trade-off when taking account of the level of risk reduction that such measures achieve. Established principles such as ALARP are in place to support such decision making when considering the proportionality of the costs associated with risk reduction options. As we start to consider systems that use AI, the number of concerns is broadened. In particular, ethical considerations come much more into focus as increasing intelligence and autonomy challenge the established assumptions on things like responsibility, accountability, reciprocity and dignity. This gives rise to additional ethical trade-offs that must be understood, justified and accepted by the affected stakeholders.

There are many proposed ethical principles. We focus on five ethical principles of non-maleficence, beneficence, respect for human autonomy, transparency and justice:

- **Non-maleficence** (the avoidance of harm) is the ethical principle that most closely aligns to our traditional view of safety. The scope of harm that is considered as part of the safety assurance process has typically been limited to physical harm to humans; the introduction of AI encourages a broader consideration of harm to include psychological harm, as well as harm to the environment and broader society. A particular issue might relate to harms that only arise in the long-term, for example through continued work monitoring systems rather than undertaking more rewarding operational or control actions.
- **Beneficence** is the ethical principle of doing good. The use of AI offers the potential for huge benefits in terms of performance and safety, especially if operators can be removed from harm's way altogether, and the ethical implications of choosing not to embrace these benefits must be considered alongside the risks associated with adoption and deployment. In general, there will be economic benefits to consider, but we exclude those from our considerations in this report.
- **Respect for human autonomy** is crucial for safety in order to ensure effective oversight. Despite increasing autonomy through the adoption of AI, humans will still play a critical role in risk control, although this role may shift from one of operator to a more supervisory role. It is essential that the responsibilities, capabilities and accountabilities of the human align to ensure that they can perform these roles effectively, and that they do not end up as a "liability sink" responsible when things go wrong, but without the time or feasible means to intervene, i.e. they have no meaningful control.

- **Transparency** is particularly important since AI is often inherently opaque, in the sense that the AI models that are used are normally not amenable to human scrutiny or analysis. This is a major concern for safety assurance which relies on evidence about how the AI is created (such as an understanding of the datasets that are used or the implementation decisions that are made). Furthermore, transparency supports explainability of AI outputs and the ability to communicate to stakeholders about critical AI properties.
- **Justice** considers how the effects of the use of AI-enabled robotics are distributed across the affected stakeholders. In terms of non-maleficence and beneficence, this requires a consideration of who benefits and who bears the risks from the use of AI. In particular we must be aware of the inevitable transfer of risk between the system and the human, and amongst different human stakeholders. A consideration of overall risk is insufficient in this regard, since an inequitable distribution of risk to a particular stakeholder group could be unacceptable even if the overall risk associated with the system operation is reduced. The extent to which the risk burden taken on by particular stakeholders is consensual is also important. Ultimately the principle of justice requires an explicit consideration of the trade-offs to ensure that the distribution of benefit, tolerable residual risk, and tolerable constraint on human autonomy is equitable across all affected stakeholders.

The AI ethics argument reasons about these principles in order to justify the safe use of the AI-enabled robotics an industrial context. The Principles-based Ethics assurance (PRAISE) framework[1] provides a basis for this argument.

Examples of AI Ethics Arguments for AI-enabled Robotics

Rather than fleshing out all the elements of PRAISE, we illustrate some of the concerns that might arise and would need to be addressed when using PRAISE for the ethics module of the BIG argument, in three particular cases. we also use these examples later in this report, and each of the three cases starts off with a brief overview of the system and setting.

Autonomous Vehicle in a Factory

FIGURE 2 AN AUTONOMOUS VEHICLE IN A FACTORY (SOURCE: MICROSOFT)

Consider an autonomous vehicle (AV), replacing a manually operated forklift truck, intended to transport finished parts in a factory from a roller conveyor at the end of the production line to another conveyor near the loading bay where they will be packaged and loaded onto trucks for delivery. Due to the factory layout, the AVs need to run parallel to the production line, consisting of a

[1] Porter, Z., Habli, I., McDermid, J. and Kaas, M., 2024. A principles-based ethics assurance argument pattern for AI and autonomous systems. *AI and Ethics*, *4*(2), pp.593-616.

sequence of machine tools, some of which are automated and some of which are operated manually; there is a defined path for the robot delineated by painted lines on the floor. The operators must also cross the AV's path to reach their workstation, e.g. for rest breaks and at the end of their shifts. The AVs have a computer vision system to enable them to detect and avoid obstacles, including machine tool operators, and track the path.

- Non-maleficence - the AV must avoid physical harms, e.g. impacting workers on the production line. It must also avoid psychological harm, alarming operators by either not giving way to them when crossing its path, only "giving way" at the last minute, or travelling close together to make crossing the AV's path difficult.
- Beneficence - reduction of the need for operators to conduct dull tasks which also carry the risk of physical injury, e.g. from vibration of the fork-lift truck.
- Respect for human autonomy - maintenance workers must be able to ensure that the AVs will not operate if they have to attend to repair a fault or to realign the AV if it is got outside its approved operating area (and halted); allowing production line operators to "pause" the AVs whilst they crossed the AV's path would also give them more autonomy and address some of the above hazards.
- Transparency - evidence that the training data for the object detection system covers the type of machine tools found in the factory (and is updated if they change), the sort of clothing worn by the operators and the lighting levels seen in the factory.
- Justice - the issues here will be twofold: First, does the use of AVs make the risks to the operators worse than when there were manually operated forklifts? Second, is the transfer of risks to the maintenance staff acceptable? This will inevitably involve comparison with the previous setup and may well involve considerations such as whether the risks are "globally at least equivalent" (GALE) as well as ALARP.

From a safety perspective, requirements will "flow-down" to the system level, e.g.

➔ AV-SR1 - detect obstacles in or approaching the AVs path at sufficient distance to stop under normal braking at least 2 metres from the obstacle/predicted nearest point of approach.

➔ AV-SR2 - maintain separation from the preceding AV so that an operator walking at typical speed (1.2 m/s) has a 5 second "window" to cross between the AVs

Robot Arm

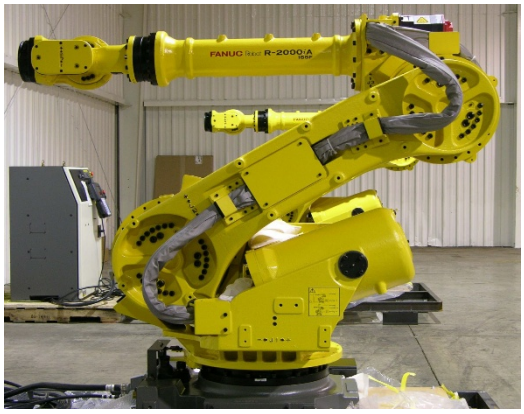


FIGURE 3 A ROBOT ARM (SOURCE: MICROSOFT)

Consider a robot arm used to support electric arc welding, acting with an expert welder, effectively operating as a cobot giving flexibility to rapidly change operational tasks in a welding cell and freeing the operator from repetitive tasks. The robotic arm has several operating modes including:

- Training - guided by the expert through a series of runs to “learn” the welding task (this will likely employ reinforcement learning).
- Optimisation - refining the learnt movements to provide a smoother trajectory thus improving weld quality.
- Operation - repeating the learnt task on presented work pieces which are loaded into the welding bay by the operator who manually initiates the welding task once the piece is in place.
- Servicing - replacement of consumables, e.g. welding rods and flux.

In addition, there may be maintenance activities. These are not considered in detail here, but similar concerns would arise as with the AV, although the set of hazards would be somewhat different.

- Non-maleficence - the robot arm must avoid injury to the operator which could arise from it applying excessive force whilst in training mode, commencing welding when the operator is in too close proximity in operation mode (as well as burns, this might cause eye damage if the operator is not wearing protective goggles when the welding operation starts).
- Beneficence - giving the operator a more varied and interesting job by automating repeated tasks and freeing the operator to work on a more varied set of tasks.
- Respect for human autonomy - the operator has the freedom (within the factory guidelines) to choose which tasks to automate, and which tasks to undertake manually.
- Transparency - this primarily relates to evidence of the accuracy of the learning and the changes in trajectory that can be introduced through optimisation.
- Justice - as with the AVs, the issues are around change in risk/benefit distribution; this should be justifiable given the nature of the capability being considered, but there are options, see below.

Note that, there are also issues to do with servicing and maintenance, but these are unlikely to be substantially different than the risks with a more conventional welding cell.

From a safety perspective, requirements will “flow-down” to the system level, e.g.

- ➔ RA-SR1 - minimise force feedback during training, e.g. to 10Nm torque, to avoid the capacity to cause harm.
- ➔ RA-SR2 - provide an interlock, e.g. via a light fence, so that the robot arm cannot enter operation mode unless the operator is at least a metre away from the work piece (note that this does not control the eye damage hazard).

There are several design options which we can consider from a justice perspective. If the welding cell has a rotating fixing assembly, so the operator loads work pieces on one side whilst the robot welds on the other side, then the assembly rotates at the end of the operation so the operator can remove the completed work piece and mount the new one, then (suitably designed) this controls the hazard related to burns without need for a light fence, but it introduces another to do with moving so that the operator’s hands get trapped. Of course, this can be controlled by requiring the operator to initiate rotation. These would be considered normal engineering trade-offs, but they can also be reflected in justice arguments considering the different risk/benefit distributions, and the level of human autonomy for each option, helping to arrive at a preferred option, ideally at concept design phase.

Humanoid Robot

FIGURE 4 HUMANOID ROBOTS (SOURCE: MICROSOFT)

Consider a humanoid robot used to provide flexible manufacturing and maintenance services in a factory which carries out machining on toxic or hazardous materials such as beryllium, cadmium or nickel alloys. To avoid hazards to human operators, from airborne dust and contaminated metalworking fluids, all the machining operations - loading the workpiece, tightening clamps, monitoring progress, removing work pieces and transferring them to the next processing stage are done by robots. Given the variety of the tasks these are done by a humanoid robot, operated through verbal commands, and providing visual feedback to the remote operator as well as using cameras itself. When the robot needs maintenance, it is intended to exit to a support area, through a decontamination facility.

The humanoid robot is designed to be operated by voice commands, with an AI system based on large language models (LLMs). The AI system contains a speech-to-text component, and the LLM operates on the resultant commands, carrying out the requested tasks, and providing verbal feedback to the operator through a text-to-speech interface. The robot comes with general capabilities which are then particularised to the factory, e.g. teaching the robot about the factory layout, the specific mounting systems on each machine, etc. The robot has multi-lingual capabilities, so it can be adapted to work in the language appropriate to the factory location/operator preferences. The robot also supports a video-streaming interface so that the remote operator has a view of the work being undertaken by the robot at any time so they can halt operations, if necessary, by issuing either a verbal command to the robot, or using a remote stop on one of the machines.

- Non-maleficence - as humans are excluded from the work area under normal operations, the risks primarily relate to physical damage to the robot or the machines; maintenance work on the robot will only be hazardous should the decontamination be inadequate.
- Beneficence - removal of operators from harm's way/ enabling operators to work without the need to wear full-body personal protective equipment (PPE) for extended periods.
- Respect for human autonomy - humans remain in control of the robot, but supervising the robot might be “demanding” in that requires a high level of attention at all time, thus giving the operator little respite; further, there is a risk that operator skills, especially those requiring manual dexterity, will erode over time.
- Transparency - a key issue will be to ensure that the localisation training means that the robot will operate “appropriately” for the particular environment, including understanding the working of each machine so that, for example, it will correctly tighten clamps for a particular model of milling machine; it is also important to be able to demonstrate that it is working in the appropriate

language at any point, especially if factory staffing means that the operational language might need to change between shifts.

- Justice - whilst the use of the humanoid reduces risk to human operators, it may put the remote operator in a situation where they are a “liability sink” - held responsible for damage caused by the robot or, for example, by it failing to tighten up a clamp properly when it is very difficult to identify such problems remotely. This problematic as it is hard to maintain vigilance overseeing operations rather than doing them, and particularly so when fast reactions might be needed to prevent a problem from escalating into a serious incident.

From a safety perspective, requirements will “flow-down” to the system level, e.g.

➔ HR-SR1 - correctly mount workpieces and fix securely prior to initiating machining.
➔ HR-SR2 - provide warnings to the remote operator if a task may not have been completed successfully, e.g. clamps aren’t successfully tightened.

Showing that these requirements have been met will require reasoning about the use of LLMs.

3. AI System Safety Argument

In any AI-enabled robotic system, AI will only ever be a small part of a larger complex robotic system consisting of a number of interacting hardware and software elements. When using AI in safety-critical systems it is therefore crucial to understand and manage the contribution of the AI to safety risk. This therefore requires that a systems perspective is taken considering in detail the operating context of the system in operation, the role of humans during the operation of the system, the organisational setting for the deployment of the system, as well as the architecture and design of the system. Established system safety engineering approaches can still be used throughout the development of an AI-enabled robotic system, and it is crucial that these proven techniques and methods are adopted as the foundation for safety assurance. Where AI is used as part of the system this brings novel challenges that may require the adoption of new approaches, or modifications to these existing practices, in order to ensure these are sufficiently addressed. In particular, AI introduces the following areas of additional uncertainty that must be addressed in the safety case.

Operating context

An AI-enabled robot must be able to respond autonomously to all situations with which it is faced. This involves continuously understanding the state of the operating environment and, based upon that understanding, making safe decisions on what actions to take. Where humans control the system, the human's innate ability to contextualise and make sense of what is happening in the environment, and to improvise when faced with unforeseen situations, often plays an important role in maintaining the safety of the system. In contrast, AI systems must be trained to deal with all scenarios that may be encountered during operation. This demands a more complete and detailed model of the operating context for the system, where gaps between the model of the operating domain and what the system encounters during operation could be hazardous. This means that the more complex and dynamic the operating environment of the AI-enabled robot is, the more challenging it is to ensure it is sufficiently modelled. It is crucial therefore to demonstrate that the operating context of the AI system is completely and correctly defined.

Hazardous scenarios identification

Hazardous scenarios are those scenarios that the system may encounter during its operation that could, under certain conditions, lead to an unsafe outcome. For AI-enabled robots, it is particularly important to identify hazardous scenarios that may arise due to the decisions that are made by the AI system in response to interactions between the system and its operating environment. This requires consideration of the belief state of the system (the understanding that the AI has of the state of the operating environment at the point at which the decision is made), along with the possible actions that the robot could take, in order to identify how hazardous outcomes could occur. It must be demonstrated that all the hazardous scenarios associated with the operation of the AI system are identified.

Safe operating concept

The safe operating concept (SOC) specifies how an AI-enabled robot must behave in order to provide a sufficient mitigation for the hazardous events that have been identified. As well as specifying required behaviour of the system under normal operating conditions, the SOC may also specify reduced operating domains (RODs) which provide additional constraints on the permitted operating context of the system under certain conditions where this is necessary to acceptably reduce the risk. A ROD may need to be defined, for example, in response to a particular system or component failure mode (such as loss of a sensor). The SOC may also specify conditions under which the capability of the AI must be reduced in order to maintain safety. It must be demonstrated that the SOC that is specified for the system is sufficient to mitigate all of the identified hazardous scenarios.

Safety requirements

As for all safety-related systems, it is imperative for AI-enabled robots that safety requirements are specified that are sufficient to ensure the fulfilment of the SOC, and that the safety requirements are correctly refined to reflect the design of the system. Safety requirements must be allocated to individual system components for implementation and verification. It is crucial for safety assurance that the intent of the SOC is maintained throughout the refinement of the safety requirements. For AI-enabled systems, the interpretation of real-world safety requirements to a form that is meaningful in an AI context can be particularly challenging, requiring a consideration of factors such as AI performance and robustness, as well as sufficiency of training data. It must be demonstrated that the safety requirements for AI are valid and traceable to the SOC and the operating context of the system.

AI system design

The design choices that are made for an AI-enabled robot will impact the ability to assure the safety of the system. In particular the chosen system architecture can be crucial to ensuring that safety requirements can be met and that hazardous failures can be mitigated. The increased uncertainty inherent in AI systems necessitates in particular consideration of the following characteristics:

- **Robustness** is the provision of correct behaviour in implicitly-defined adverse situations arising due to an uncertain system environment. Robustness is therefore a characteristic that is particularly important for AI-enabled systems since it helps to mitigate hazardous system failures associated with hard to predict, and thus unexpected, changes in a complex operational environment.
- **Fault-tolerance** is the provision of correct service despite faults arising from the system itself. The focus of fault tolerance is therefore on the detection and recovery from anticipated failures identified as part of the hazard assessment process.
- **Runtime monitoring** enables the behaviour of the AI system during operation to be checked against defined constraints or behavioural predictions. The runtime monitor should be independent from the components it is monitoring but may take the same inputs. One of the biggest challenges with using runtime monitors is being able to correctly define the constraints or bounds on behaviour that the monitor will check. It is also desirable to monitor behavioural trends of the AI system post-deployment to identify longer-term changes in performance. It must

be demonstrated that the design decisions that are taken for the AI-enabled robot are appropriate to ensure the safety requirements can be met throughout operation.

Hazardous failures management

All systems are susceptible to hazardous failures of the system components or of their interactions. For AI-enabled systems, there are a range of novel failure modes that must be considered and analysed and for which mitigations must be determined. The specific failure modes can often not be identified until the details of the design solution and in particular the characteristics of the system components is understood. One method for determining potentially hazardous failures within the system is to consider possible deviations from intended behaviour that may arise during the operation of the system. For example, industrial robots may use AI components for perception tasks (to identify objects in their environment). The types of failures that should be considered in this case include[2]:

- Unrecognised sensor failure – Failure, degradation or malfunctioning of sensor or sensor data is not recognised
- Not detected – Objects appear in the sensor field of view but are not detected
- Not classified – Objects are detected but are not classified and their presence is therefore rejected
- Misclassification – Objects are detected but are misclassified, e.g. as being static when they are mobile leading to a failure to predict motion.

It must be demonstrated that all potentially hazardous failures within the system are identified and acceptably managed.

Out of context operation

We have already discussed the importance of defining the operating context for the AI-enabled robot. This represents the scope within which safe operation can be assured. The nature of AI-enabled systems means, however, that no matter how well we undertake this task it is almost inevitable that the system will spend some time operating outside of that defined context, perhaps due to unanticipated changes in the operating environment. When this happens the behaviour of the robot could be unsafe. There are several situations where operation outside the defined context may occur for the AI-enabled system:

1. The environment or context of the AS suddenly changes without warning, for example a flood in a warehouse may not have been predicted as part of the operating context for the system. When changes such as this arise suddenly and unexpectedly it is not possible for the robot to anticipate and avoid these conditions.
2. The robot fails to recognise the boundary of the defined operating context so is unaware that it is operating in potentially unsafe conditions. This could be due to the boundary being poorly defined or ambiguous or limitations in the sensing capability of the robot.

[2] Molloy, J. and McDermid, JA., 2022. Safety Assessment for Autonomous Systems' Perception Capabilities. *arXiv preprint arXiv:2208.08237*.

3. The robot may recognise the boundary has been reached but is unable to hand over safely to another function or an operator (either because none are available or the transition itself fails). In such situations the safest option may be for the robot to continue to operate. It may also take time for the robot to recognise that is has failed to make the transition.

It is important therefore to assure that the boundary of the safe operating context can be recognised by the AI-enabled robot, accepting that in some cases it may not be possible to directly detect this using the sensors available to the robot. In such cases it may be that "proxy" measurements are required to be used. It is then necessary to determine a strategy for what the robot should do to minimise the risk if it finds itself operating out of context. The minimum risk strategy will vary depending upon the particular system and its operation. For example, in many cases the safest strategy may simply be for the robot to return to a situation where it is within the safe operating context as quickly as possible. In many cases however, this may be difficult or even impossible for the robot to achieve. It is possible that a set of principles or heuristics may need to be employed to cover all situations. Where it is possible, the minimum risk strategy may include the system handing control over to a human operator.

Verification

In many regards, the approaches to verification of AI-enabled systems are no different from those used for traditional systems performing safety-related tasks. However, as the behaviour of AI is often less well bounded than for traditional systems, this can give rise to particular challenges, particularly when verifying systems operating in complex environments. These challenges include[3]:

1. Unpredictable Environment: This unpredictability adds uncertainties to testing as the system can encounter environmental variables that may have been unknown during design.
2. System and Scenario Complexity: It is unclear how to define scenarios that include all features involved in the operational environment and the system itself and how to check whether the scenarios used for testing reflect the actual situations encountered.
3. Data Accessibility: To be able to test the systems, good quality data is required. This data can be difficult and costly to collect, interpret and validate.
4. Missing Standards and Guidelines: Standards and guidelines are generally not established for testing of AI-enabled system. This makes determining the sufficiency of the testing activities used more difficult, and harder to justify.

It is crucial that an appropriate verification strategy is justified and adopted. This should consider the use of both testing and formal verification techniques. It is likely that for AI-enabled robots significant use may be made of simulation as part of the testing process. Simulation may be used for a number of reasons: it can allow verification activities to begin earlier in the development lifecycle, since a fully developed system is not required; it can also enable tests to be carried out that would otherwise be difficult to create (due to their rarity in the real world) or dangerous to perform (since they would require

[3] See: Qunying Song, Emelie Engström, and Per Runeson. Concepts in testing of autonomous systems: Academic literature and industry practice. In 2021 IEEE/ACM 1st Workshop on AI Engineering- Software Engineering for AI (WAIN), pages 74–81. IEEE, 2021.

exposing people to unnecessary risk); it can also be used as a way of increasing the coverage that can be achieved through testing, by enabling test cases to be created quickly and cheaply. The disadvantage of using simulation is that all simulations are models of the real world and justifying the representativeness of the simulation models can be challenging.

It must be demonstrated that the verification undertaken is sufficient to demonstrate that the defined safety requirements are met throughout the whole of the defined operating context (and out of context). The SACE framework[4] (Safety of Autonomous Systems in Complex Environments) provides a structure for the AI system safety argument that specifically addresses each of the AI challenges described above.

Examples of AI System Safety Arguments for AI-enabled Robotics

The discussion here continues with the three examples introduced when considering the ethics argument. For brevity, it focuses on the SOC and verification, with a particular emphasis on the identified DSRs and ways in which hazardous failures or out of context operation could contribute to hazards (violate the derived safety requirements). The second and third examples are discussed in less detail as many of the general points, e.g. about real-world testing versus simulation, and the role of a verification plan are generic, and discussion the AV is sufficiently general to cover the other two cases.

Autonomous Vehicle in a Factory

The safety requirements identified in the ethics argument were:

- ➔ AV-SR1 - detect obstacles in or approaching the AVs path at sufficient distance to stop under normal braking at least 2 metres from the obstacle/predicted nearest point of approach.
- ➔ AV-SR2 - maintain separation from the preceding AV so that an operator walking at typical speed (1.2 m/s) has a 5 second “window” to cross between the AVs

There are, in effect, functional requirements which will form part of the SOC. They also need to be implemented and thus will “flow down” through the design decomposition (the safety requirements stage). The SOC, however, needs to be defined to ensure that the system is robust, including covering hazardous failures and out of context operation. Assuming that computer vision is used to detect obstacles and the lines delineating the AV’s path, then we can add additional DSRs:

- ➔ AV-SR3 – monitor state of the cameras, and trigger a controlled stop on the current trajectory, on the occurrence of pre-defined failure modes (links to AV-SR1 and AV-SR2).
- ➔ AV-SR4 – monitor trajectory of the AV with respect to the edge of the path, and if it will exit the path at the current steering angle and speed, intervene to:
    1. Correct the course, bringing the AV back onto the path, or:
    2. If 1 is not successful, bring the AV to a stop, minimising the excursion from the path

[4] Hawkins, R., Osborne, M., Parsons, M., Nicholson, M., McDermid, JA. and Habli, I., 2022. Guidance on the safety assurance of autonomous systems in complex environments (SACE). *arXiv preprint arXiv:2208.00853*.

AV-SR4 is essentially "new" and relates to out of context operation. Both AV-SR3 and AV-SR4 should be considered as potential hazardous scenarios and checked against other requirements to remove or to resolve conflicts. In practice, this might need to be referred to the ethical level if, for example, there was a conflict between requirements to stay in the path and to avoid impact with an operator.

When the SOC is first defined, manual analysis or simulation-based testing should be used to explore the DSRs to see if it is reasonable to expect them to be satisfied given assumptions about the implementation[5]. It is likely that simulation would be most effective, but it is important that the simulation reflect the AV dynamics, the uncertainty of the operational environment, e.g. variation in wheel-floor friction caused by spilt oil or freshly painted lines, and a realistic range of plant layouts.

Towards the end of the development, once the lower-level components have been implemented and individually verified, system-level verification will be carried out to confirm overall system behaviour. This would include end-to-end testing, for example to see whether the response time in implementing AV-SR4 is sufficiently short to avoid impact with machinery, etc. It is likely that some of this will be done in simulation, and some with physical testing. It is expected that the approach would be defined in a verification plan (extending normal practice to AI/autonomy) with considerations including the risk of doing the test in the real-world and the fidelity of the simulation (the better it is, or the better its limitations are understood) the more can be done effectively and adequately in simulation.

Robot Arm

The safety requirements identified in the ethics argument were:

- ➔ RA-SR1 - minimise force feedback during training, e.g. to 10Nm torque, to avoid the capacity to cause harm.
- ➔ RA-SR2 - provide an interlock, e.g. via a light fence, so that the robot arm cannot enter operation mode unless the operator is at least a metre away from the work piece and terminates operation if the operator breaches the light fence (note that this does not control the eye damage hazard).

The SOC, as well as incorporating these top-level requirements, needs to be defined to ensure that the system is robust, including covering hazardous failures. We can illustrate two further DSRs:

- ➔ RA-SR3 – monitor the torque being generated and progressively reduce it to zero[6], and terminate the training session, providing an audible alarm to the operator (links to RA-SR1).
- ➔ RA-SR4 – monitor the light fence and shut down the welding operation (turn off the current) and move the welding head a small distance from the workpiece so that it does not attach to the workpiece (links to RA-SR2).

These are both hazardous failures. The notion of out of context operation is less obvious, although it might include loading an inappropriate workpiece, i.e. one on which the robot hadn't been trained, and initiating operation as opposed to training mode. If this was deemed credible then thought would need

[5] Some would refer to this as validation, not verification.

[6] The rate of reduction should be confirmed with ergonomists, as a very sudden reduction might lead to harm to the operator if he or she is applying a lot of force (torque).

to be given to implementation mechanisms – a form of RFID tag or bar code might be possible, but there is then an issue of ensuring that the right tag or bar code is affixed. For the RA, it might be possible to do more of the initial and final verification with the physical system rather than simulation as a lot can be done to ensure safety during verification by disabling the welder (turning off the current).

Humanoid Robot

The safety requirements identified in the ethics argument were:

- ➔ HR-SR1 - correctly mount workpieces and fix securely prior to initiating machining.
- ➔ HR-SR2 - provide warnings to the remote operator if a task may not have been completed successfully, e.g. clamps aren't successfully tightened.

As with the other two examples, the SOC needs to ensure robustness to hazardous failures and out of context operation. Unlike the other two examples, however, we consider implementation issues, e.g. the use of LLMs and image analysis, in considering hazardous failures.

- ➔ HR-SR3 – initiate and execute a dialogue with the operator, confirming (links to HR-SR1).:
  1. the appropriate machine for the workpiece;
  2. the workpiece orientation if there is more than one option, and
  3. "talk through" the fastening process.
- ➔ HR-SR4 – when providing warnings to the remote operator if a task may not have been completed successfully, e.g. clamps aren't successfully tightened, ensure that the video stream shows the issue, e.g. the clamp in question (links to HR-SR2).

These two DSRs are slightly more subtle than the earlier examples. With HR-SR3, the sub-cases relate to the (potential) hazardous failures:

1. LLM misinterpreting the name/number of the machine.
2. LLM misinterpreting the nature/design of the part.
3. LLM (likely reinforcement learning process) carrying out what is known as "reward hacking" getting to the end-state efficiently, by skipping some of the intermediate steps.

HR-SR4 is intended to mitigate the case where the robot correctly points out its error, but shows an image which doesn't include the error, misleading the operator to over-ride the warning, assuming the robot got it wrong. NB this may be a case where the operator becomes a "liability sink". In this case it might not be possible to do much (useful) system-level verification prior to implementation.

4. Purpose-Specific AI Model Safety Argument

In this section we consider how the safety of the AI components of the system can be assured. In particular we focus on the use of Machine Learning (ML) to a provide specific AI functionality as part of the system. When used to perform specific tasks such as image classification, ML has been shown to often achieve exceptional levels of performance. Safety however is rarely concerned just with overall performance; the ML component must behave in a safe manner in all situations it may encounter, and we must be able to demonstrate that this is the case. For industrial robots that are required to operate in complex, open and dynamic environments this is particularly challenging since the operational space is so large and uncertain. It is crucial therefore that safety assurance of ML is framed within the context of the overall system, and particularly the system safety requirements that the ML must satisfy if it is to be considered safe for its particular purpose within the defined system context. Crucial to this is traceability between the technical claims regarding the AI component and the higher-level safety claims at the system and ethical/societal levels. It is necessary to justify how the system-level safety requirements are broken down into specific technical AI safety requirements, and to generate evidence to demonstrate the requirements have been satisfied. This requires consideration of safety assurance throughout the entire ML lifecycle. The following stages of the ML safety assurance process are defined in the AMLAS methodology[7]. For each stage there are a set of activities that must be undertaken in order to generate the evidence required to support the AI system safety case.

ML Safety Assurance Scoping

Given that safety claims about ML can only ever be made with respect to the context of the system in which ML will be used, the role ML plays within that system, and the environment in which the system operates, it is critical to explicitly define the scope under which we are able to demonstrate the safety of the ML component. Here we can link to information created as part of the AI system safety argument to ensure integration of the overall safety case.

ML Safety Requirements Assurance

The safety requirements identified at the system level capture complex real-world concepts and are not of a suitable form for ML development. It is therefore necessary to translate these into requirements which can be used in the construction and verification of an ML model. One of the primary foci of this stage is to justify the sufficiency with which the derived ML safety requirements capture the intent of the specified in the safety requirements. As a minimum that the set of ML safety requirements must include requirements on both the performance and robustness of the ML model (where model robustness considers the model's ability to perform well when the inputs encountered during operation are different to those present in the training data). Robustness requirements must encapsulate features of the operational context of the system that have been identified as having an effect on the model output, such as environmental features like light levels and types of objects, or sensor noise.

[7] https://www.york.ac.uk/assuring-autonomy/guidance/amlas/amlas-pdf/

Data Management Assurance

The data used to train ML components directly impacts their performance. Therefore, the development and assurance of data requirements is essential for safety. Data requirements for an ML component should include consideration of data relevance, accuracy, balance, and completeness. **Data relevance** refers to the extent to which the development data is representative of the operating environment and architecture of the system into which the ML component will be deployed. **Data accuracy** considers the extent to which variations in data gathering, pre-processing and labelling can impact the satisfaction of the ML safety requirements. **Data balance** typically considers the number of samples for each class present in the data sets. Ideally all data sets used would be perfectly balanced, i.e. the same number of samples would exist for every class of interest. In practice however those samples which are of particular interest in a safety context are often, by their nature, more difficult to obtain. While balance considers the number of samples for each class, **data completeness** concerns how the collected data sets reflect the robustness requirements specified in the ML safety requirements. This will consider the extent to which all the identified features of concern in the operating context of the AI-enabled system are present in the data samples. Ensuring completeness is challenging in complex industrial environments in which the combinations of relevant features become large. Having defined the data requirements it must be demonstrated that ML data which meets these requirements is obtained and validated.

Model Learning Assurance

The creation of ML models is a highly iterative process with numerous decisions points concerning things such as model structure, learning strategy and hyper-parameter selection. Each of these decision points and the supporting rationale for decisions should be recorded and justified. ML design decisions will not be purely driven by safety but will also take account of these wider factors such as performance and reusability. It should be shown therefore that these decisions do not compromise our ability to satisfy the safety requirements.

Model Verification Assurance

Evidence must be generated to demonstrated that the ML model that has been created satisfies the defined ML safety requirements when exposed to inputs not present during the development of the model. Verification is typical undertaken through testing of the model using test data sets. Verification data should be deliberately challenging for the ML model whilst also realistic (within the defined operating context of the system). Verification data should therefore be gathered using an adversarial mindset and by people who are independent of ML model development to avoid “training to the test”.

Model Deployment Assurance

Once the safety of the ML model is assured, it is also necessary to consider the safe integration of that ML model into the target system. Up to this point, the AMLAS process has focused on the assurance of the ML model itself. This integration process typically involves attaching the components to its inputs, such as the robot’s sensors and pre-processing pipelines, as well as connecting the outputs to traditional software units which ultimately produce outputs to the robot actuators. A key part of this integration is

to identify what the deployment assumptions are that could impact the safety of the ML component if they are violated. Assumptions should consider the hardware upon which the model will execute, the nature of the system into which the ML component will be integrated (such as the type of sensors that are used); and supporting software libraries which may be different to that used at development time.

It is important to also consider the assurance of the ML component through-life as it is used in evolving open industrial environments. This requires the creation of monitors that can detect when the ML model is being used outside of its defined safe scope.

Examples of Purpose-Specific AI Model Safety Arguments for AI-enabled Robotics

For this stage we consider verification and deployment generically then cover robustness aspects of safety requirements and the completeness criterion from data management for the first two examples The humanoid robot may well contain some purpose-specific AI, but we do not discuss it here, instead focusing on the use of LLMs in the next section where we consider general-purpose AI.

Verification will likely involve a mixture of real-world and simulation (synthetic environment) testing. Simulation is good for exploring a wide range of situations, for example factory layouts for the AV, especially if the simulation can be run faster than real time. It is also possible to use simulation to explore behaviour in challenging (hazardous) scenarios or failure conditions, for example by modifying sensor data to represent the effect of glare from lights (perhaps on another AV). However, real-world testing is needed for at least two reasons. First, to know that the simulation has sufficient fidelity that its results can be treated as adequate for assurance purposes (they don't need to be "exact", e.g. photo-realistic images in simulations may not be necessary). Second, to assess behaviours that are hard to assess with confidence in simulation, e.g. what the coefficient of friction is between the AV wheel and newly painted lines on the floor for the path boundary and how this affects the stopping distance. The choice of approach should be documented in a verification plan, and the justification for defining the plan should be included in the overall safety case (maybe just a summary from the plan).

Deployment is concerned with the use of the AI in the physical system in which it is embedded. This will include sensor location (does the training data sufficiently reflect the position?), the vibration induced by movement of the robot (is this covered in robustness testing?), and so on. Also, work needs to be done to address variability between individual system builds or in mounting sensors and actuators, e.g. the robot arm, to see if the variations are within limits assumed in developing the system. It should also be noted that such things can change with time, e.g. vibration increasing due to flats on tyres following an emergency stop. Depending on the impact of such deployment issues, there may be a need for maintenance action to replace the tyres, so the AI algorithms continue to operate correctly. It is likely that this would be considered in a verification plan, with a rationale for the balance between generic testing and the testing of individual units (perhaps factory acceptance testing).

Autonomous Vehicle in a Factory

In principle, the safety requirements should be refined from the system level to the level of the ML model, for example AV-SR1 might be refined to the level of the number pixels in the image needed to detect an object at suitable range, this is outlined below, with parameters that would need to be determined from analysis of the cameras and the AV's braking system. We also add a robustness requirement that needs to parameterised according to the operating environment, which is assumed to include the use of strobe lights on AVs to help make them more visible.

- ➔ AV-SR1-refined - detect solid obstacles that present as [XX] pixels or more in an image and classify them as "non-driveable" with at least 90% confidence, and with a range of geometries representative of real-world objects.
- ➔ AV-SR1-robustness1– meet AV-SR1-refined at least [PP] frames out of [QQ] when subject to strobe lighting of intensity [LL] Lux for [SS] milli-seconds.
- ➔ AV-SR1-robustness2 – do not classify as "non-driveable". objects that meet the requirements of AV-SR1-refined but will not impede the AV and will not be damaged by the AV driving over them.

The 90% confidence may seem low, but the assumption is that the AV software will "integrate" over several frames before taking an action, e.g. emergency braking, and this is realistic for computer vision systems based on neural networks. The first robustness requirement reflects an understanding that the strobes on other AVs might "blind" the cameras for a short period as they rotate, and that AV-SR1-refined still needs to be met. The second robustness requirement reflects the fact that there might need to be cables across the path, in protective trunking, if temporary power is needed, e.g. for maintenance work. This shouldn't lead to the AV stopping. Of course, these requirements could be merged, but they are kept separate as there are likely to be other robustness requirements and the combined requirement would be very complex.

The requirements for data coverage can easily be illustrated considering AV-SR1-refined. The images used to train the AI need to represent the types of machines likely to be in the factory. They should cover all angles of approach to the machine, i.e. 360 degrees in the horizontal plane, and be representative of the field of view given the mounting height of the camera. Further, the data should cover both images of obstacles where the AV should stop, and those where it shouldn't – see AV-SR1-robustness2. Note that, if the camera mounting is changed then the models might need retraining.

Verification should demonstrate that these requirements have been met. As noted above, this is likely to require a mixture of testing in simulation and the real world. For example, it is likely that AV-SR1-robustness1 is best addressed by real-world testing as simulating the effect of strobes on cameras is likely to be difficult to do with sufficient fidelity.

Robot Arm

This is rather different to the AV case, as the system uses AI in the training phase, and the resultant algorithms in the operation phase. It is not realistic to expect an assurance argument to be generated each time the arm is trained – and a different strategy is needed.

As specified, it is realistic to meet key requirements from the SOC using conventional software, e.g.

- ➔ RA-SR2 - provide an interlock, e.g. via a light fence, so that the robot arm cannot enter operation mode unless the operator is at least a metre away from the work piece and terminates operation if the operator breaches the light fence.
- ➔ RA-SR3 – monitor the torque being generated and progressively reduce it to zero, and terminate the training session, providing an audible alarm to the operator (links to RA-SR1).

Thus, it is possible to build the assurance case around these protection mechanisms, rather than having to directly assess the AI components in the safety case. This is a good strategy, where it is viable; in this case it is likely to be practicable. However, the concepts of robustness testing and coverage also apply to conventional software[8]. For RA-SR3 coverage needs to consider feasible trajectories for training, that is the changes in direction and angle of the welding head, the likely range of speeds for welding (likely quite narrow to retain good weld quality), and so on. In practice, the coverage should be skewed towards "high torque" actions, e.g. when going round corners on a workpiece, and away from other situations, e.g. a straight seam, where the forces are likely to be comparatively stable.

For robustness, we need to consider reasonably foreseeable adverse events to which the system must be robust. For example, for RA-SR2 the system needs to respond fast enough to limit harm (thermal effects cannot be eliminated by the machine as the welding rod and workpiece will be hot) by stopping the welding operation quickly. In this case, it is necessary to consider the maximum speed with which the operator might move through the light fence. This might be because they see a problem and unwisely move to try to stop it, or perhaps because they are moving round the workstation and trip and fall towards the machine. Similarly, perhaps the worst-case situation in respect of RA-SR3 would be an involuntary movement such as startle reflex to a sudden noise or someone approaching them unexpectedly (maybe not seen or heard if coming from behind), or a sneeze. In both these cases an analysis of realistic operating conditions is needed to understand robustness requirements – and to generate the tests to achieve coverage of those requirements.

[8] The frameworks for assuring AI build on established good practice so this is no accident.

5. General-Purpose AI Model Safety Argument

As mentioned above General-Purpose AI (GPAI) models include Large Language Models (LLMs). At their heart, LLMs predict the next word in a sequence, having been trained on a very wide corpus of data (a sanitisation of the text on the internet). This capability is not very useful, in itself, so many "tools" have been built on top of LLMs most notably Chatbots such as ChatGPT or Claude which can engage in a dialogue with a user through a query interface. There are, however, other forms of models such as Visual Language Models (VLMs) which combine image analysis with LLMs and can, for example, explain the scene in an image. Further, Visual Language Action Models (VLAMs) combine these capabilities with the ability to perform physical tasks. VLAMs seem ideal to provide the "brain" for many classes of robot, especially those that rely heavily on perception for understanding the environment around them and deciding on a course of action.

As with all AI models, one of the challenges in assurance arises from the complexity and opacity of the models. Many current GPAI systems are based on transformers, a particular class of neural network, better adapted to handling sequences of data, e.g. text or video streams. However, they are highly complex, and state-of-the-art models might contain more than a trillion parameters (internal values that govern how the model works). Even if these parameters, often called weights, could be exported for human scrutiny it is not possible to assess the behaviour (or safety) of the model by direct inspection or analysis. The approaches to the assurance problem are normally architectural, but a few observations are worth making to provide a richer context.

First, as GPAI models are general-purpose they often do remarkably well at answering questions as they will have seen something similar somewhere in their training data. However, this will not be "your" data, and the answer may not reflect the specifics of your problem. This can be addressed by techniques such as Retrieval Augmented Generation (RAG) where the model is focused on specific data in answering questions – for example, about the layout of a factory where a robot is intended to work.

Second, GPAI models suffer from hallucinations or more precisely confabulation where the model fills in gaps in their "memory" with fabricated, distorted, or misinterpreted information. This is intrinsic to the models, although there is research seeking to "tame" this tendency. It is worth noting that RAG techniques don't solve the confabulation problem – the models can still fill in gaps in the data with fabricated material.

Third, GPAI models don't have a good understanding of physics, or causation, or even basic maths (although again this is the subject of research). Thus, the models might be given a simple problem of calculating a braking distance, given parameters of speed, mass, friction, etc. – and not get it right.

Fourth, the applications built on top of the models, such as Chatbots are designed to be pleasing, so they will present positive views on information presented (this can be overcome by directing the Chatbot to be more balanced). This, in part, explains the hallucination/confabulation problem but it also exacerbates it. The outputs from such tools are designed to be compelling, thus they can seem very plausible, even if they are wrong. This makes human oversight difficult.

Pragmatically, the above issues mean that GPAI models must be treated as untrusted components of systems. This may not be a problem if hazard and safety analysis shows that the consequences of errors are insignificant. If this is not the case, i.e. the modes can affect safety, then there are essentially three different approaches that can be adopted.

First, extensive testing that builds confidence in the (use of the) models. This would allow an estimate of the unsafe failure rate to be produced and might be acceptable so long as the acceptable rate was quite high[9]. Even if this is done there would be residual uncertainties due to the complexity and opacity of the model which it would not be practicable to quantify.

Second, it is possible to use architectural means, often referred to as "guardrails" to prevent undesirable behaviour. Most obviously, these guardrails can be on the outputs from the model where they can, for example, be checked for physical feasibility or safety. Such guardrails might be technical (a piece of conventional software) or human; however, both the difficulty of humans operating in a monitoring role and the propensity for the models to be plausible even if wrong, means that this is not necessarily an effective approach. Guardrails on the inputs seek to remove data that could cause undesirable outputs. This might be "obvious" such as a command to drive at twice the speed limit, but more subtly it might be something that triggers a failure behaviour of the model, e.g. through confabulation. Thus, to provide guardrails on inputs requires an understanding of failure modes which realistically can only come from testing which will be far from exhaustive.

Third, it is possible to generate explanations of the outputs of the model to help build confidence in what it is doing. There are many possible uses of explanations[10] however the most useful in this context are pre-deployment to give general confidence in the model and in operation giving explanations of specific behaviours[11]. Although valuable, it should be noted that explanations are simplifications and approximations of what the model is doing thus they indicate the underlying logic, but don't necessarily reflect it accurately.

In practice, it is likely that a combination of methods for assuring GPAI will be chosen – although the first step should be to ask whether the task in hand is suitable for implementing using GPAI, based on an assessment of hazards and risks including how failure modes of the models might be mitigated.

GPAI itself and applications of GPAI are evolving at a bewildering pace. Thus, the above should be taken as being representative at the time of writing and likely to remain broadly accurate in terms of the basics of how the models work and problems they can exhibit. However, care should be taken to review the state-of-the-art before developing and deploying systems employing GPAI.

---

[9] To demonstrate a failure rate of 10^{-X} per hour requires more than 10^{X+1} hours of testing, even at modest confidence levels.

[10] McDermid JA, Jia Y, Porter Z, Ibrahim Habli I; Artificial intelligence explainability: the technical and ethical dimensions. *Philos Trans A Math Phys Eng Sci* 4 October 2021; 379 (2207): 20200363.

[11] See, for example, https://wayve.ai/thinking/lingo-natural-language-autonomous-driving/ which illustrates the use of explanations of the behaviour of an autonomous vehicle, which is driven using VLAMs.

Examples of General-Purpose AI Model Safety Arguments for AI-enabled Robotics

Humanoid Robot

The humanoid robot uses LLMs to conduct a dialogue with the user, including to meet requirement HS-SR3. For explanatory purposes it is assumed that use of LLMs is acceptable and the outline of an assurance approach is sketched out below, but it is left to the reader to form a judgment whether the level of assurance that can be achieved is sufficient given the attendant risks.

➔ HR-SR3 – initiate and execute a dialogue with the operator, confirming (links to HR-SR1):

1. the appropriate machine for the workpiece;
2. the workpiece orientation if there is more than one option, and
3. "talk through" the fastening process.

An architectural approach is taken, combined with systematic testing. LLMs have a broad vocabulary; that is not needed for this problem, so the first step is to introduce an input guardrail that limits the input from the operator to commands and requests relevant to the task. This is likely to be done by defining an ontology, i.e. a structured representation of the relevant knowledge, such as the classes of machine tool that are in use, the types of fasteners used, and so on. This would enable the system to accept commands such as "show the workpiece mountings" and to filter out commands such as "put the spaghetti into the lathe". This input guardrail can be specified, developed and tested stand-alone to show that it only passes through appropriate commands and queries to the LLM.

Similarly, it is possible to design guardrails on the outputs. There will be constraints on the sequence of actions that can be undertaken, e.g. the workpiece must be mounted and fastened before machining starts. Similarly, lubrication must be started before cutting starts. However, more subtle constraints may be needed, such as fastening specific clamps first before turning the workpiece to fix other clamps, where doing this in the wrong order would lead to the workpiece falling risking damage. Such ordering constraints can be specified, and an output guardrail developed that only allows through valid sequences[12]. Again, this can be tested stand-alone including checking that the output explanation (HR-SR3 3) matches the physical commands issued. The ability to reject invalid sequences can be tested by providing out of order commands or interspersing irrelevant commands in a valid sequence.

Having designed and verified guardrails, the system can be integrated and tested as a whole. The notions of coverage and robustness discussed earlier are still relevant here. Coverage would involve testing every command and perhaps every pairwise combination of command and object, including invalid combinations to check that they are rejected. Robustness would include providing a workpiece of type A and giving commands for a workpiece of type B, to check that the error is detected. It would likely also include detecting faulty workpieces, e.g. where a mounting flange didn't contain a fixing hole. Also,

[12] Li, C., Lai, P., Zhang, N. et al. EcoRxAgent: an AI agent for generating economically substitutable prescriptions. npj Digit. Med. (2026). https://doi.org/10.1038/s41746-026-02612-7 gives an example of this sort of guardrail, albeit in a healthcare context, but the principles are largely domain-independent.

robustness to variations in operator accent and background noise would be needed as the system uses a verbal (speech) interface.

A richer architecture could be envisaged that also include explanations. For example, a command could be given “explain the fastening sequence before implementing it” enabling an operator to check what was going to happen, as well as having the output guardrail. Queries could also be used “what torque did you use in fastening that bolt” to check aspects of the operation that can’t be validated simply through sequence checks by the output guardrail (the output guardrail could be made more complex but it would likely have to depend on the robot to accurately report torque values).

The introduction of guardrails, e.g. limiting the dialogue with the robot, and perhaps the use of explanations might seem to reduce the value of introducing such technology (it certainly increases the cost). However, there is still enormous potential in the use of such technology to give flexibility, even if the guardrails need to be updated to deal with new machines or new types of workpieces. This loops us back round to the ethics argument – do the benefits outweigh the risks, especially given the challenges of assurance, and is there sufficient human autonomy in controlling the system?

As a final observation, many countries, including the UK, have government-backed centres focusing on evaluation of GPAI. In the UK, the focus is more on security rather than safety, but it is worth noting that security breaches can lead to unsafe behaviours. In general, the approaches employed by these centres are based on evaluations of the models against benchmarks, i.e. using challenging test scenarios, and so-called “red teaming”, i.e. getting experts to try to “break” the system and to produce undesirable behaviour. This work might help us to understand potential failure modes of GPAI but it does little directly to provide assurance about downstream applications of GPAI, e.g. in a humanoid robot. Finding the most cost-effective way of linking the “upstream” evaluations and the “downstream” uses of GPAI remains an open research question[13].

[13] McDermid JA, Jia Y, Habli I. Upstream and downstream AI safety: Both on the same river? arXiv preprint arXiv:2501.05455. 2024 Dec 9.

6. Conclusions

This report has considered the challenges of introducing AI-based robotics into factory settings and how we might provide assurance so this can be done safely. The use of AI can introduce new forms of risk, e.g. bias, and thus the need to consider broader ethical issues, not just safety. To address this, we introduced and illustrated the "BIG Argument" which links ethical considerations to system level safety assessment and the assurance of AI, covering both conventional approaches such as neural networks and general-purpose AI such as large language models (LLMs). The report gives complete, but succinct, descriptions of these four core elements of the BIG Argument[14] and illustrates them with three simplified examples – a factory autonomous vehicle (AV), a robot arm used in welding, and a humanoid robot employed as a flexible machinery operator. These examples are all based on a requirements flow-down, showing how ethics and safety are reflected at all levels in the design. Note that we did not have space to consider assurance of changes – but that is an important issue which should be addressed in any real-world application of AI-enabled robotics.

The AV sheds light on some of the trade-offs that are necessary between different ethical considerations when there are people undertaking operational and maintenance tasks in the proximity of robots. It also illustrates direct assurance of AI used in the perception component of the system, including robustness to adverse conditions in the factory environment.

The robot arm again illustrates trade-offs but shows where it is more appropriate to gain assurance from conventional software mechanisms to "guard" the behaviour of the AI-enabled robot arm, rather than to address the AI itself. These two examples therefore illustrate two contrasting approaches to assurance of systems using special-purpose AI – each valid in the appropriate context.

The humanoid robot explores the issues of assuring LLMs used in control of a robot, and the challenges associated with general-purpose AI. There are significant difficulties in assuring GPAI directly and the most widely used assurance approaches of evaluations against benchmarks are not directly relevant to safety-related applications such as robotics. Thus, we illustrated architectural means of "guarding" the behaviour of the humanoid robot as a basis for providing assurance. It is an ethical judgment whether such assurance is sufficient for a given application.

Done well, safety assurance can be an enabler of innovation and guide the developers and deployers of new technologies towards safe and effective solutions. The approaches we have outlined here are intended to achieve that balance between assurance and innovation that allows effective but justifiable uptake of new technology. In such a fast-moving field, this report will not be the last word, but we hope it provides some useful guidance for current and emerging technologies. Finally, if we were to speculate about the medium-term evolution of these technologies, it would be that the community will find ways of harnessing visual language action models in such a way that they allow the deployment of flexible and adaptable robots, and which can also be effectively assured.

[14] Details can be found through the references.

Global Initiative
For Industrial Safety

Cambridge Industrial
Innovation Policy

UNIVERSITY OF
CAMBRIDGE

The Global Initiative for Industrial Safety

The Global Initiative for Industrial Safety (GIFIS) unites global policymakers, innovators, and industry leaders to safeguard working environments through emerging technologies. Through knowledge sharing and community-driven solutions, we reduce risks, enhance resilience, and elevate workplace safety worldwide.

www.industrialsafetyinitiative.com