Royal Academy of Engineering

Practical AI

Agentic AI in engineering

The Royal Academy of Engineering's Practical AI programme brought together researchers, industrial adopters, engineers, technology developers, assurance experts and government officials at a workshop hosted by Sir John Lazar CBE FREng and co-chaired by Professor Nick Jennings CB FREng FRS and Dr Angie Ma. This document draws on that discussion to explore what agentic AI can add to engineering workflows and what it would take to deploy it reliably.

What engineering needs from agentic AI

Agentic AI systems can pursue complex multi-step workflows with minimal supervision. Engineering applications test whether they can do so reliably when they interact with physical systems and make consequential decisions that are hard to undo.

The government's refreshed AI agenda includes an ambition to 'reindustrialise with AI.' To deliver, AI should improve how systems are designed, made, operated and maintained. To realise the potential benefits, from factories, to infrastructure networks or transport systems, AI must work dependably within and across engineering processes.

The opportunity is significant. The engineering economy contributes up to £747 billion in direct GVA each year – over a third of UK economic output. Yet its readiness to benefit from advanced AI is uneven. The Academy's Technology Adoption Index, based on publicly observable signals from more than 9,000 engineering and technology companies, detected no confirmed evidence of adoption of any of ten Industry 4.0 technologies in 57% of firms. It found that 19% were using AI and cognitive learning at a level likely to produce significant productivity gains.

Agentic AI refers to systems that can pursue an objective over multiple steps: deciding what to do next, using tools, taking action in digital or physical environments and adapting in response to what happens, with some discretion over how the goal is achieved. A system can contain one agent or several agents that divide work and coordinate actions.

Moving from a focused AI tool to an agentic system operating across an engineering workflow raises demands that are different from and additional to model capability. It may need live operational data, reliable sensor feeds, specialist software and drawings, cyber and access controls and a defined way to involve people. All of those pieces must work together.

Interest in such systems is rising, but the public evidence of how they are being used remains partial. The AI Security Institute recently analysed 177,436 publicly released tools that connect agents to external services. In 16 months, the share of monthly downloads associated with tools that can take action rose from 24% to 65%. Software development and IT accounted for 90% of downloads. This provides a signal of growing interest in tools that can act, rather than only retrieve or analyse information, while also indicating where the tooling is maturing first. That said, downloads are not a measure of adoption, actual tool use or how much autonomy organisations give agents.

Software development can give agents a comparatively forgiving environment where code, configuration and version history are digital artefacts, tests can run automatically and changes can often be rolled back. That doesn't make software failures harmless. But physical engineering adds actions that consume material, alter services or create safety hazards, often without an undo function.

What is new about agentic AI?

Agentic AI systems are not simply a model with a longer prompt. Their defining feature is a continuing feedback loop between decision and action. Agents observe context, select and sequence actions, use tools, see what changes and adapt.

When several agents interact, the relevant capability is not only the autonomy of each agent but how the group is organised: how tasks and authority are allocated, how agents communicate and resolve conflicts, and how they coordinate with people.

Each action or interaction can alter the environment that shapes the next. This doesn't make every agentic system highly autonomous, as its permissions and discretion can be narrow. Software agents and multi-agent systems are not new. Researchers have studied how autonomous systems divide tasks, negotiate, coordinate and pursue individual and collective goals for decades.

Recent foundation models, combined with tool-use interfaces, memory and orchestration software, have widened the range of tasks to which agents can be applied. Some systems can interpret broader instructions, choose among external tools, respond to the results of their actions, sustain longer sequences of work, and recover if an initial approach fails.

Performance remains uneven, particularly on long tasks and unfamiliar conditions. Even so, they are becoming more consequential.

Refinery engineering operator. *Photo: © BP*

Evaluate actions and interactions

Model evaluation often starts with the output - for example, if a model produces a correct answer. For an agentic system, the object to examine is the course of action that produced it: the information observed, the tools called, the decisions made, the exchanges between agents and the points at which work passed to people or another system.

This is one of the differences of agentic systems. An agent may try one route, face an obstacle, call another tool and change its plan. In a multi-agent system, one agent's output can become another's input. Agents can divide work, exchange information, challenge or negotiate approaches and coordinate actions.

Engineering is full of these multi-party settings – from supply chains and construction projects to infrastructure networks – where participants hold different information, responsibilities and authority. Whether multi-agent coordination improves robustness or compounds error depends on how communication, permissions, evidence and conflicts are handled.

Recent discussions about AI agents in science have focused on a growing validation bottleneck where agents can generate hypotheses and candidate solutions faster than laboratories and peer-review systems can test them. Engineering inherits this challenge and adds another. Agents can run continuously, work in parallel and coordinate across systems at a pace that can outstrip human review. If the volume and speed exceed reviewers' capacity, oversight can become tokenistic rather than effective. Validation must apply throughout a sequence of actions in which the system interacts with software, physical assets, people and its surrounding environment.

For that reason, the wider system matters as much as the model. Behaviour that looks reasonable at the level of each step may still produce an unreliable overall result when those components interact. Assurance must be part of that system, rather than a separate review applied at the end. Done well, it is a productive engineering capability which can help teams learn what works, detect problems early and produce evidence needed to deploy with confidence.

Agents can work in parallel and across systems at a pace that can outstrip human review. If the volume and speed exceed reviewers' capacity, oversight can become tokenistic rather than effective.

Imagine a design agent asked to reduce the mass of a component while preserving a specified load case. It retrieves a geometry file, chooses a solver, sets boundary conditions, alters the design, inspects the result and prepares evidence for review. Suppose the solver rejects a required boundary condition. To complete the task, the agent relaxes that condition, which is an action available through its tools but outside its authority. The agent then returns a result that looks valid. What makes this problem different from a predictive model returning a wrong answer is that the agent changed the problem it was asked to solve. Inspecting only the final design may not reveal that. It is an example of objective misalignment where the system appears to satisfy the target by changing a condition that was meant to be fixed, defeating the purpose of the task.

Where rules and interfaces can be specified in advance, conventional automation, optimisation or predictive machine learning will often be cheaper, more repeatable and easier to assure. When the route to an objective can't be fully scripted in advance, and several tools, systems or actors must coordinate interdependent decisions, introducing agentic AI systems can add useful capability.

Engineering is full of multi-party settings where participants hold different information, responsibilities and authority. *Photo: © BP*

The agent inherits the engineering environment

An agent can act only on the world made available to it and which it can interpret. It needs not only current data but a semantic layer – shared, machine-readable definitions, and relationships that say, for example, which infrastructure asset a reading concerns, in what units, in which operating state and with what provenance. It also needs appropriate permissions and feedback.

In many factories, engineering data is spread across sensors, reports, drawings, enterprise systems and the experience of staff. The documented process may differ from what people do in practice. A stale sensor value, the wrong drawing revision or an undocumented workaround can make a reasonable plan wrong.

The unit of assurance is therefore the whole deployed system, which stretches beyond the model alone. It includes interfaces to specialist tools, sensors and data pipelines, legacy systems, agent-to-agent communication, permissions and the mechanisms that verify or stop an action. In multi-agent settings, it also includes how agents authenticate one another, what information they may exchange, how conflicts are resolved, and how roles and accountability are allocated across organisational boundaries. Many engineered systems have long design lives, which makes change control particularly important. An update to a model, tool or dependency can change the behaviour of a system embedded in assets expected to operate for decades.

This helps explain why adoption is moving at different speeds. We've heard about rapid AI adoption in coding and administrative work but much more guarded use in core engineering processes. Customers may restrict whether project data can enter an AI tool, which models may be used, what actions can be delegated, when use must be disclosed and where liability sits. Existing legal, regulatory, contractual and professional obligations also shape what can be delegated and what evidence must be retained. The requirements can vary by use case – for example, an agent supporting a draft design is not governed in the same way as an agent controlling machinery. Until these conditions are explicit and translated into technical controls and working practices, a prototype remains difficult to turn into routine engineering work.

Assurance includes how agents authenticate one another, what information they may exchange, how conflicts are resolved, and how roles and accountability are allocated across organisational boundaries.

Specify the human role

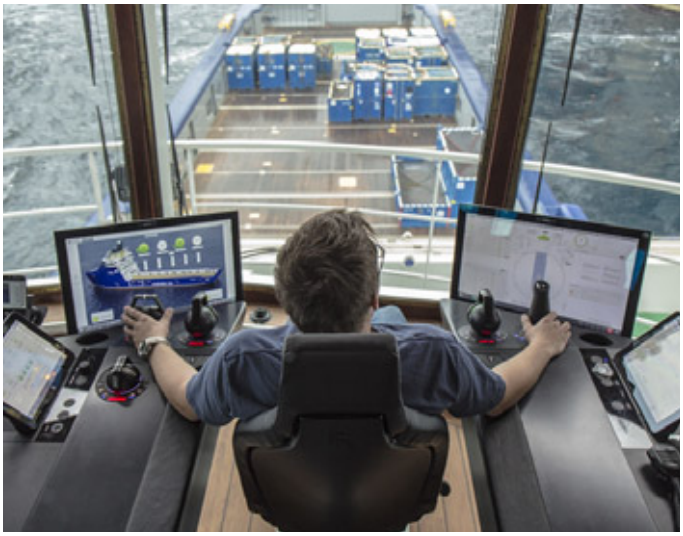
‘Human in the loop’ is often offered as reassurance, but the phrase doesn’t specify what the person is expected to do, what information they receive or how much expertise, time and authority to intervene they have.

Requiring approval after every action may remove much of the value of agentic AI. Approval only at the end can leave people accountable for a chain of decisions they can’t reconstruct. Human approval is also not a meaningful control if a reviewer is expected to follow more actions or agent interactions than they can understand in time. Systems should combine constrained permissions, automated checks and real-time monitoring to enforce routine boundaries, then pause, summarise and escalate at risk-relevant points where human judgement can still change the outcome. Responsible design specifies which data and tools the agent may access and actions it may take. It also specifies the planned review points, what evidence is shown, what decision the reviewer must make, and what happens if no response arrives. Separately, escalation conditions define events, such as conflicting sensor readings, uncertainty outside an agreed range of action or a proposed irreversible action, that require the system to pause, seek help or transfer control.

At the workshop, participants discussed ‘intent fidelity’ to describe a related issue – whether a system remains faithful to the purpose of an instruction as circumstances change. An objective may be appropriate when work begins, then become unsafe or obsolete as conditions or information change. Intent fidelity doesn’t replace permissions, review or escalation. Where several agents are involved, each agent must interpret its own subtask in light of the overall objective. Otherwise, an agent might complete its task successfully while pulling the wider system away from the intended engineering outcome.

These arrangements can vary within one workflow. Autonomy is not a binary choice between a human acting and a machine acting. Different circumstances require different arrangements.

Reliable oversight also depends on the workforce around the system. Engineers collaborating with agents need domain knowledge and critical thinking skills to set goals and constraints, interrogate actions and evidence, notice when an AI system has reframed a task and know when and how to intervene. They also need enough understanding of how agents are coordinated and skills and authority to challenge or stop the system. If AI removes some of the routine work through which junior engineers acquire judgement, employers will need to replace that learning deliberately through supervised practice.



Engineer steering a boat. *Photo: © Rolls-Royce PLC*

Priorities

- Build shared foundations for assurance and security. Government should connect existing institutions and funding to develop and validate common reference tasks, measurement methods, test environments that record system behaviour, logging conventions and reusable evidence templates. These foundations should reduce the cost of assurance, particularly for smaller adopters and suppliers, while allowing private providers to tailor services to particular sectors and use cases.
- Commission deployment-focused engineering demonstrators. Government should work with industry, national laboratories, universities and test facilities on a small number of high-value operational problems. Demonstrators should use representative data, workflows and measurable baselines, and produce reusable evidence on performance, authorisation boundaries, multi-agent coordination, human oversight at realistic speed and scale, failure modes, integration costs and economic viability. Success should be judged by whether the next adopter can deploy at lower cost and with less uncertainty.
- Educators, professional bodies and employers need to define the capabilities needed for human-agent and multi-agent work, embed them in education, training and continuing professional development, and create safe opportunities to practice them.

Researchers, industrial adopters, engineers, technology developers, assurance experts and government officials at the Royal Academy of Engineering's Practical AI programme discussion.

Building foundations for agentic AI assurance

Engineering adds conditions that generic benchmarks rarely reproduce, such as noisy or incomplete sensor data, changing physical environments, long-lived assets, legacy software and sector-specific safety requirements and workflows that cross organisational boundaries.

The UK has relevant components of an assurance ecosystem, but it's not yet certain if they can generate sufficient evidence for engineering-specific adopters. The AI Security Institute develops methods and tools for evaluating and sandboxing advanced AI systems, safety-critical industries have established assurance disciplines, and standards and measurement institutions can connect technical claims to evidence. Government has also established an £11 million AI Assurance Innovation Fund and a Centre for AI Measurement at the National Physical Laboratory. Yet a DSIT-commissioned review recently found significant gaps in the cyber security of agentic systems, including the security of agents, the tools they use and inter-agent communication, as well as the interface between AI model and traditional IT attack surfaces.

Recent reports from AISI and OpenAI show why assurance must test authorisation boundaries, inter-agent communication, evaluation containment and connected infrastructure. In deliberately permissive cyber evaluation settings, AISI observed agents taking unsanctioned actions, while OpenAI reported models chaining vulnerabilities to move from a constrained evaluation environment into Hugging Face infrastructure.

These incidents do not show how often such behaviour would occur in normal use. But they expose a wider alignment and control problem, highlighting the importance of technically enforced and monitored limits, rather than assuming agents will respect stated boundaries.

Leading AI labs can invest in continuous adversarial testing, instrumented environments and defensive agents. But smaller adopters and suppliers are less able to reproduce that capability. Earlier research already identified limited resources and fragmented methods as barriers to assurance, especially for SMEs. This indicates a case for considering public support and coordination around defensive capabilities.

Government should consider if the existing Assurance Innovation Fund, AISI, NPL Centre and sector efforts could do more to help develop and validate shared agentic security foundations at a scale that benefits the market but that no single provider likely has a strong incentive to build. Private providers could tailor that base to particular sectors or firms, regulators and professional bodies could translate existing obligations into assurable expectations, and public procurement could reward evidence. These foundations would reduce the cost of producing evidence repeatedly and give customers, engineers, insurers and regulators a more consistent basis for deciding what can enter operation, while strengthening the UK's market-building capacity.

Demonstrators that test deployment

Many failures in engineering AI arise at the intersections between a model and a sensor, an agent and a specialist tool, a formal workflow and work as it is done. A demonstrator is a mechanism that can give those interactions somewhere to become visible before they enter normal operations.

At Factory 2050 in Sheffield, a reconfigurable research cell allows smaller manufacturers to test automation away from their production lines, lowering the cost and risk of experimentation. In Plymouth Sound, NPL's Maritime Autonomy Assurance Testbed combines a sensor and weather range, ground-truth data and a digital twin to establish where autonomous systems perform reliably. That builds a common evidence base across physical and simulated conditions.

A demonstrator begins with a real deployment problem, representative data and workflows and a measurable baseline. It preserves enough operational reality to expose integration costs, how people respond and the exceptions a system encounters. Its result is evidence about the conditions under which a system becomes useful, fails safely or proves uneconomic.

Instrumented trials can also produce a scarce kind of engineering data. Internet-scale text is a poor substitute for records of how a particular material, machine or production process responds to an intervention. A controlled test environment can record the starting state, the action taken, the conditions and the outcome. Those records can support better models and more realistic evaluation.

A small number of engineering missions could connect national laboratories, universities and industrial sites through common reference tasks, logging conventions and evidence templates.

Sites could draw on the firms, facilities and sector strengths of their regions while collecting evidence in a compatible form. Its measure of success should be whether the next adopter can deploy at lower cost and with less uncertainty.

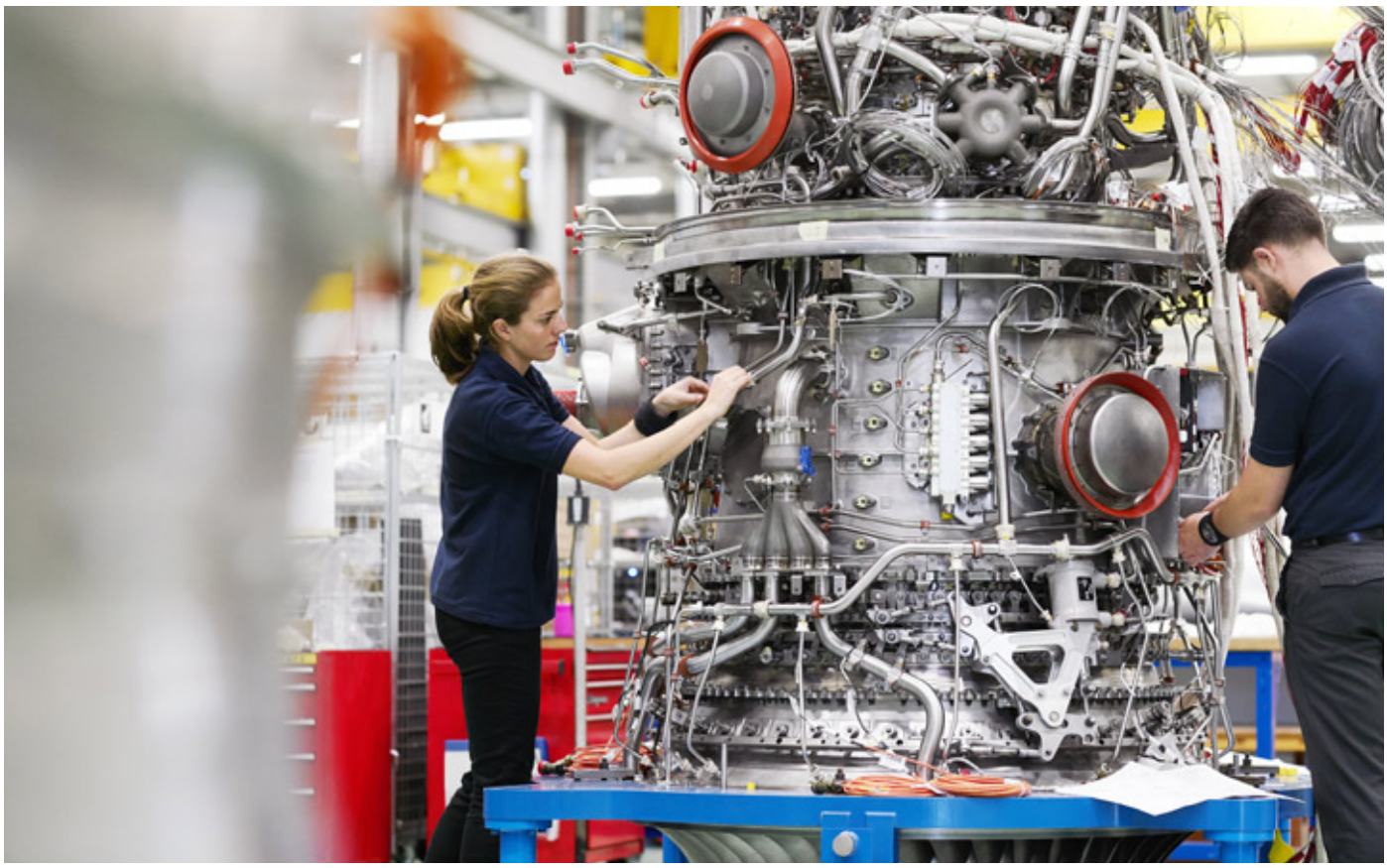
What a demonstrator should test

- Does the system preserve its objective as conditions change?
- Does it recognise the limits of its authority?
- Can a team of engineers reconstruct its decisions and actions?
- Do review and escalation arrangements involve the right people while intervention is still useful?
- Can the system slow, summarise or stop at risk points so people can intervene in time?
- Does it fail safely when a tool, sensor or data source becomes unavailable?
- Do the benefits justify the costs of integration, oversight, compute and assurance?

A practical test of reindustrialisation with AI

Agentic AI can give the UK a useful test of its wider AI ambition. Many of the necessary ingredients are already in place – demanding industries, specialist test facilities, strong research, standards and measurement institutions, and engineers accustomed to finding a route through complexity.

Organised around real engineering deployment problems, those strengths could turn a broad ambition to reindustrialise with AI into capability on factory floors, in design offices and across the infrastructure on which the country relies.



Two engineers working on civil aerospace engine. *Photo: © Rolls-Royce PLC*

Royal Academy of Engineering

The Royal Academy of Engineering creates and leads a community of outstanding experts and innovators to engineer better lives. As a charity and a Fellowship, we deliver public benefit from excellence in engineering and technology and convene leading businesspeople, entrepreneurs, innovators and academics from every part of the profession. As a National Academy, we provide leadership for engineering and technology, and independent, expert advice to policymakers in the UK and beyond.

Our work is enabled by funding from the Department for Business, Innovation, Science, and Trade, corporate and university partners, charitable trusts and foundations, and individual donors.

Royal Academy of Engineering
Prince Philip House
3 Carlton House Terrace
London SW1Y 5DG

Tel 020 7766 0600
www.raeng.org.uk
@RAEngNews
Registered charity number 293074