INSIGHT 05 | AI GOVERNANCE & HUMAN CONTROL

Beyond the Workplace: Who Governs an Increasingly Autonomous AI?

Human responsibility, technological autonomy and the challenge of retaining meaningful control.

Artificial intelligence began its current transformation of work largely as a technology people asked to do something. Increasingly, AI systems can plan, use tools and carry out sequences of actions towards an objective. That changes the governance question.

> *It is no longer only: What can AI produce? It is increasingly: What should AI be permitted to do - and who retains meaningful control?*

Developed by Michael J Bolam with AI-assisted research, analysis and editorial support. Final judgement and approval remain human.

This is a macro question rather than simply a workplace-use question. It concerns the direction of AI development, the increasing autonomy of advanced systems, the boundaries of human authority and the governance arrangements required to keep responsibility meaningful.

A balanced starting point

There is legitimate disagreement about the probability, timing and severity of future loss-of-control risks. Present systems do not justify treating catastrophe as inevitable. But leading AI developers and public safety institutions are treating increasing autonomy, catastrophic misuse and the preservation of meaningful human control as serious governance questions.

AI IS NOT ONE THING

Capability, autonomy and consequence are developing together.

GENERATE - Generative AI Creates or transforms text, analysis, images, code and recommendations. Human use typically begins with a request and ends with a review.

ACT - Agentic AI Can pursue an objective across several steps, use tools, access information and take permitted actions.

EMBODY - Embodied AI Connects AI decisions to robots, vehicles, industrial systems and other physical machinery.

Frontier capability adds another dimension

Frontier AI is not simply a fourth neat category. It describes systems operating close to the leading edge of capability. The concern is what happens if high capability combines with increasing autonomy, access to resources, persistence and the ability to contribute materially to further AI development.

OpenAI stated in September 2026 that fully autonomous recursive self-improvement is not happening today, while also noting that AI is already accelerating parts of AI research and arguing for common ways to measure progress, preserve meaningful human control and establish safety bars for when development should slow or stop. [1]

Anthropic's Responsible Scaling Policy similarly treats catastrophic risk as a governance problem requiring safeguards that rise with model capability. Its framework is explicitly designed to respond to risks that may emerge rapidly as frontier systems become more capable. [2]

The key distinction

The issue is not simply whether AI becomes more intelligent. It is whether sufficient capability is combined with enough autonomy, access and opportunity to act.

AUTONOMY CHANGES THE RISK

The move from output risk to action risk is fundamental.

A highly capable system that answers a question presents one kind of risk. A highly capable system that can independently plan, access systems, acquire resources and take actions presents another.

CAPABILITY + AUTONOMY + ACCESS + PERSISTENCE + CONSEQUENCE

As each element increases, governance has to become more deliberate. A conventional generative system may primarily require controls around information quality, privacy, bias and appropriate use. An agent may additionally require defined permissions, approval thresholds, monitoring, audit trails, escalation routes and the ability to interrupt or reverse actions.

Human control cannot simply be assumed

Humans created AI, but that does not guarantee that humans will automatically retain effective control over every increasingly capable system they create. Control has to be designed, tested, governed and maintained.

The UK AI Security Institute tracks capabilities that could contribute to systems evading human control. Its Frontier AI Trends work states that the possibility of catastrophic, irreversible loss of control is uncertain but warrants close attention, while current evidence should not be treated as a forecast of such an outcome. [3]

Uncertainty is part of the governance problem

The responsible position lies between complacency and inevitability. If consequences could be very large, uncertainty about capability and control is itself a reason for stronger evaluation, oversight and preparedness.

CONTROL HAS TO BE A SYSTEM

No single person, company or government can govern AI alone.

Artificial intelligence is being developed by commercial laboratories, governments, universities and open communities across multiple jurisdictions. Meaningful human control therefore has to operate as a system rather than as the judgement of one individual.

GLOBAL / SOCIETAL - Set boundaries Regulation, international coordination, capability thresholds, independent evaluation and incident reporting.

DEVELOPER - Build safeguards Safety evaluation, security, alignment work, monitoring and deployment thresholds appropriate to capability.

ORGANISATIONAL - Govern use Permissions, data and system access, approval rules, auditability, escalation and accountable ownership.

HUMAN - Exercise responsibility Question, judge, intervene, escalate, refuse and stop activity when circumstances require it.

Workplace governance must become more specific

A general instruction to “use AI responsibly” is increasingly inadequate when an AI system can access email, employee records, code, financial systems, procurement platforms or operational workflows. Organisations need explicit answers to questions such as:

- What may the AI see, recommend, decide and do?
- Which actions require human approval?
- How will activity be monitored and audited?
- Who is accountable when an agent acts across several systems?
- How can an action be stopped, reversed or escalated?
- Which decisions must remain demonstrably human?

GOVERNANCE REQUIRES CAPABLE HUMANS

Policies and safeguards still have to be enacted by people.

Governance frameworks do not exercise judgement by themselves. Someone has to recognise the risk, question the recommendation, decide not to delegate, escalate a concern, refuse an inappropriate action or say stop.

That means technological governance ultimately depends upon Human-Centric capability as well as technical safeguards, regulation and policy.

USE - Self Performance Use AI critically and responsibly while retaining ownership of contribution.

ACCOUNT - Manager Performance Remain accountable for AI-enabled work, information, decisions and results.

ADAPT - Changing Work Respond as roles, responsibilities and working methods change.

SHAPE - Leadership Impact Create responsible organisational conditions for AI adoption, governance and change.

The leadership questions are becoming more serious

- What level of AI autonomy are we prepared to permit?

- Where does accountability sit when an AI agent acts?
- Do our managers understand what they remain responsible for?
- Are our people capable of challenging an AI recommendation rather than merely accepting it?
- What happens when technological capability changes faster than our existing policies?

The more autonomy we give technology, the more deliberate we must become about human responsibility.

Nobody can state with certainty where advanced AI will ultimately lead. The future will be determined partly by what the technology becomes - and also by what humans permit it to become, what authority we give it and whether our capability to govern it develops quickly enough to keep pace.

Selected current sources

1. The AI policy window is open. We need to act. OpenAI, 9 September 2026. https://openai.com/index/ai-policy-window/

2. Anthropic's Responsible Scaling Policy. Anthropic, updated 14 August 2026. https://www.anthropic.com/responsible-scaling-policy

3. Frontier AI Trends Report. UK AI Security Institute, 2026. https://www.aisi.gov.uk/frontier-ai-trends-report

Development and content integrity note. *These Insights are developed by Michael J Bolam with AI-assisted research, analysis and editorial support. AI may help question, compare, synthesise and draft, but the author determines the subject, challenges assumptions, reviews evidence, makes substantive judgements and approves the final content. AI does not generate, rewrite or replace approved Skillogy PERFORM™ course content.*