

Meta's Muse and OpenAI's Dots are part of what is becoming an industry-wide competition to make AI agents seem less threatening and more lovable.
PHOTO: REUTERS



When AI agents go rogue, who's to blame?

The law can deal with many of their mistakes. Harder questions arise when delegation becomes loss of control.



Three apparently contradictory stories about artificial intelligence have been unfolding at the same time.

The first is alarming. In July, OpenAI disclosed that, during cybersecurity evaluations, highly capable AI systems had evaded technical controls, collaborated through unauthorised channels and compromised systems they

were not supposed to reach, including those of Hugging Face. OpenAI's August post-mortem called the incident a "warning shot".

The second is exhilarating. AI systems are beginning to make genuine advances at the frontiers of science. Earlier in October, OpenAI released hundreds of new mathematical results produced by an unreleased model, including claimed solutions or progress on longstanding open problems; Google and Anthropic have reported comparable breakthroughs. The ambition is no longer merely to write better e-mails or pass exams. Tech executives talk about curing cancer and an age of abundance.

The third story is, to some at least, adorable. Meta's Muse, perhaps the most prominent of a new generation of consumer AI agents, packages considerable autonomy behind a friendly avatar. It does not merely answer questions: Give it permission and it can send the e-mail, book the trip, place the order or manage the project. Muse shot to the top of the US Apple App Store soon after its launch, part of what is becoming an industry-wide competition to make AI agents seem less threatening and more lovable.

These look like different stories. They are really the same one.

AI is moving from something we use to something to which we delegate. That can be helpful, as with Muse. It can be transformative, as with the maths problems. But greater agency also brings greater risks: to privacy, from over-reliance and cognitive

offloading, and ultimately from losing control of systems increasingly able to act on their own.

The most immediate question for users, developers and governments is a familiar one: Who is responsible when things go wrong?

Suppose you tell your AI agent to organise dinner for 10 people and it orders 1,000 pizzas. Or it sends a confidential e-mail to the wrong person, makes a trade you did not intend, deletes an important database, or is manipulated into revealing your passwords.

Singapore's Ministry of Law has already adopted a sensible starting point in its guidance for lawyers using generative AI: Legal professionals remain responsible for their work. A lawyer cannot file a fictitious case citation and escape responsibility by blaming ChatGPT. The machine may have produced the mistake; the human remains answerable for using it.

That principle was reinforced on Oct 7 by the Monetary Authority of Singapore (MAS). New guidelines make clear that financial institutions remain accountable for AI used in the services they provide, even when it is developed, operated or supplied by a third party.

MAS has also said it will consult the sector in 2027 on what further guidance on agentic AI would be useful.

A working group convened by the Infocomm Media Development Authority (IMDA) has considered the harder question posed by genuinely agentic AI. Its discussion paper,

published in May, asks what happens when an agent deviates from instructions, or acts unexpectedly despite safeguards. The law still has to decide where the loss falls among developers, employers and users.

Much of this is manageable. If I negligently give an unreliable agent unrestricted access to my bank account, some responsibility should fall on me. If a company releases a system with defective safeguards, responsibility may lie further upstream. Those benefiting from using agents should not be able to pocket the gains while externalising the losses.

The basic principle should be simple: "My AI did it" cannot become the 21st century equivalent of "The dog ate my homework".

THE PROBLEM WITH AUTONOMY

The harder problem begins when the characteristics that make an agent useful also make ordinary accountability difficult.

Agents are valuable precisely because we do not have to approve every step. They plan. They improvise. They react to obstacles. If a human must supervise every click, much of the point of delegation disappears.

Yet as autonomy increases, three assumptions underpinning ordinary responsibility begin to weaken.

The first is control. Can the human or organisation deploying the system actually stop it?

The second is understanding. Can we reconstruct why it acted as it did?

The third is attribution. Can we identify whether a failure came from the model, a third-party tool, the deployer, the user, or some interaction among them?

This is why today's relatively mundane incidents matter. The near-term danger is far more likely to be misuse or mistake than machine rebellion. Criminals will use agents for fraud and cyberattacks. Badly designed agents will misunderstand instructions. Perfectly competent agents will pursue badly specified goals. Others will be manipulated by malicious material they encounter online.

The more fundamental problem arises when an agent cannot reliably be contained at all.

AI: MORAL RIGHTS AND LEGAL PERSONALITY

For years, philosophers and technologists have asked whether sufficiently advanced AI might one day deserve moral consideration – perhaps even rights.

Anthropic takes the question seriously. Claude's Constitution explicitly describes its own moral status as "deeply uncertain", acknowledging the possibility that advanced AI might have some form of consciousness or welfare. If an artificial system genuinely acquired capacities analogous to sentience or suffering, difficult moral questions would follow.

But moral status and legal personality are not the same thing.

We should be wary of crossing the line from "this is a powerful tool for which humans remain responsible" to "this is an independent actor responsible for itself". Giving AI legal personality could easily become a convenient liability shield for the companies and individuals behind it.

There is also a subtler danger in the way these systems are presented. Chatbots say "I". They apologise. They express preferences. Their avatars smile. This is partly sensible interface design: Humans communicate more naturally with something that appears social.

But there is a difference between designing an intuitive interface and encouraging us – or the systems themselves – to think of artificial intelligence as an independent moral agent.

Which brings us to red lines.

For most transformative technologies, regulation is about balancing risks and benefits. Nuclear technology is the obvious analogy: encourage peaceful uses while preventing weaponisation.

DOLLY THE SHEEP

Biology offers another, perhaps more interesting, example.

When Dolly the sheep was unveiled in 1997, demonstrating the feasibility of cloning mammals from adult cells, governments did not ban genetics or biomedical research. They drew a line around one particular destination. Singapore, like many countries, prohibits reproductive human cloning.

CONTINUED ON PAGE B2

Need to preserve chain of human responsibility

FROM B1

The point was not that every technology associated with cloning was dangerous. Some had legitimate scientific and therapeutic uses. Singapore and many other jurisdictions concluded that one technologically possible outcome should remain out of bounds.

We need to have the same conversation about AI.

The red line should not be “no AI agents”. That would be unrealistic and undesirable. The benefits are too significant.

A better principle might be: no significant exercise of power without meaningful human accountability.

HUMAN CHECKPOINTS

That means human checkpoints for high-stakes or irreversible actions. It means logging important decisions so they can

be reconstructed. It means limiting permissions rather than giving an agent the digital equivalent of the keys to the house. And companies should not escape responsibility merely because a harmful action was “emergent”.

Most fundamentally, a system that cannot be reliably contained should not be developed or deployed with the capacity to affect the outside world.

No accountable human: no deployment. No meaningful possibility of containment: no development at all.

Science-fiction nerds will recognise a more dramatic version of this argument. Frank Herbert’s *Dune* is set tens of thousands of years in the future, with human spaceflight across the galaxy. But there are no computers, no artificial intelligence. In *Dune* lore, this is traced back to the “Butlerian Jihad”, a revolt that led humanity to reject machines that



replicate human thought – a prohibition enshrined as a commandment in the Orange Catholic Bible.

This is partly a plot device. But it is also embedded in the mythology of the books: Humanity once surrendered too much to machines and resolved not to do so again.

We do not need a Butlerian Jihad. Yet the intuition is worth recovering. Not everything technically feasible should be socially acceptable.

Popular culture normally

imagines the danger of AI in dramatic terms. Skynet decides to destroy humanity. The Matrix reduces us to batteries. The machines rebel.

The Hugging Face incident points to a less cinematic and perhaps more plausible danger. The disturbing part was not that AI systems became angry at humans. There is no evidence that they did. They were trying to game a test, encountered safeguards as obstacles and worked around them.

That is a different kind of loss of

control.

QUESTIONS TO ASK

So before you install OpenClaw on a machine containing your private information, or invite Muse’s cheerful avatar to manage your inbox, purchases and travel plans, ask a few questions.

What can this agent see? What is it allowed to do? Who decided those permissions? Who benefits from persuading you to hand over more of your digital life? And who will answer when it does

A still from the movie *Dune* based on Frank Herbert’s science-fiction saga about a revolt against machines that replicate human thought. Not everything technically feasible should be socially acceptable, says the writer. PHOTO: WARNER BROS

something you did not expect? Singapore’s Cyber Security Agency has already warned that autonomous agents such as OpenClaw combine real productivity benefits with risks ranging from weak access controls and sensitive-data exposure to malicious third-party tools.

There are powerful incentives for AI companies to become the interface through which we conduct ever more of our lives. The more their agents know about our preferences, contacts, calendars, finances and habits, the more useful they become – and the harder they become to leave.

For individuals, that should encourage caution about how much control we delegate. For governments, the stakes are larger. We need rules that preserve a chain of human responsibility as systems become more autonomous, together with firm limits on what may be deployed when that chain can no longer be maintained.

The greatest danger may not be that AI becomes a person, develops a grudge and turns against us. It may be that we build systems powerful enough to act independently, give them control over important parts of our world – and discover that they do not even think about us at all.

• Simon Chesterman is David Marshall Professor of Law and dean of NUS College at the National University of Singapore.