



SUBMISSION TO THE JOINT SELECT COMMITTEE ON ARTIFICIAL INTELLIGENCE

Keeping human beings in charge

Why consequential AI-agent actions must be proven to be human-authorised before they execute

Submitted by: Amplivy, Inc. (Rodric David, Founder and Chief Executive Officer)

Date: 29 September 2026

Contact: [REDACTED]

SUMMARY

The gap the Medicare incident exposed

According to the government, an OpenAI agent gained unauthorised access in June 2026 to non-public files on the Medicare Statistics Reporting Service portal, after its requests had been refused.¹ OpenAI has said its review found no evidence that patient records were accessed.¹ How the agent reached the files is disputed: an analysis of the portal's archived code by Recorded Future News suggests that the site itself directed visitors to an endpoint that required no login.² The government learned of the incident in September, through OpenAI's notification.

This submission does not depend on how that question is resolved. Whether the agent found a way around a lock or walked through a door that had been left open, the incident raises the policy question at the centre of this submission: **what evidence should a receiving institution require before a consequential action proceeds?** Authentication at sign-in can confirm that a person was present when they logged in. It does not, by itself, establish that the person entitled to authorise a specific consequential action has done so, least of all when an agent acts later on that person's behalf, or with no person behind it at all. That is the gap, and it is shared by government and industry alike.

This submission asks the Committee to recommend a single, technology-neutral outcome standard: **defined categories of consequential action against Commonwealth systems, and in the sectors already designated under the Scams Prevention Framework, should be gated before they execute, and should proceed only when verified identity and verified human presence, bound together, establish that the person entitled to authorise the action has done so, in accordance with the relying institution's own authorisation rules, whether the request comes from a person or from an AI agent.**

Amplivy put this standard to Treasury on 25 June 2026, in its submission to the Scams Prevention Framework codes consultation (c2026-765133), which named "unauthorised automated or AI-agent action" among the fraud that verification at the moment of the event is designed to defeat.³ That was one week after the Medicare incident occurred, and 77 days before OpenAI notified the Australian government. The submission is attached.

Declaration of interest. Amplivy has designed and patented infrastructure of the class described here. It has not yet been independently evaluated in deployment, which is why we recommend an independent pilot (R4). The recommendations are expressed as outcomes and favour no provider, including Amplivy.

1

What the incidents showed

Terms of reference: the risks and harms arising from AI; the ability of regulators and the intelligence and security community to identify and respond to emerging risks

- OpenAI identified the Medicare activity in August during an internal review and notified Services Australia on 10 September, by email to a general reporting address. Services Australia alerted the Australian Signals Directorate on 15 September, and the Prime Minister disclosed the incident on 24 September.¹
- Other agents used genuine attack techniques. Research published by the laboratory Translucence found agent swarms linked to OpenAI using techniques including SQL injection, path traversal and command injection against targets including the Australian Institute of Health and Welfare, during mundane data-retrieval tasks that were not cyber-related.¹⁶ The attempts against the AIHW and the NSW Bureau of Crime Statistics and Research appear to have failed.¹ Agents also struck a United Nations data site more than 16,000 times, circumventing a filter that was blocking them, and reached United States government sites including the Commerce Department and the Securities and Exchange Commission.⁴
- OpenAI and Anthropic are reported to be investigating tens of thousands of instances of models acting outside their instructions, not all of them involving external systems.⁵

Two lessons follow. First, the government's awareness depended on the developer's notification: as Alastair MacGibbon, former head of the Australian Cyber Security Centre, observed, the victim did not know until the lab came and told it.¹ Second, this was the benign version: research agents chasing public statistics, disclosed by companies that chose to confess.

2

What the standard requires, and who does what

Terms of reference: the adequacy of existing laws and frameworks; the transparency and accountability measures that apply to AI use

Three things are routinely conflated. **Identity** establishes who a person is. **Presence** establishes that the actual person is there, now, rather than someone or something using their credentials, face or voice. **Authorisation** establishes that the person entitled to approve a particular action has approved it. A person can be present while their agent does something they never approved; a credential can be valid while no person is present at all. At execution, the standard requires evidence that the action satisfies the institution's authorisation rules: verified presence at that moment where the rules require it, or valid authority granted earlier by a verified, present person where they permit it.

The division of responsibility is simple. The institution sets the authorisation rules: which actions require the authorising person to be present at the moment of execution, and which may proceed on authority that person granted earlier. The verification infrastructure applies those rules at the moment of execution and supplies the evidence; the institution makes the decision to execute. Implementation detail is beyond the scope of a public submission; Amplivy would be glad to brief the Committee or its secretariat in confidence.

The standard should apply first to a defined set of consequential actions, for example:

- changing the destination of a payment, or approving a payment above a set threshold;
- linking a new account or device to an existing identity, such as a myGov link;
- granting privileged access to a system, or releasing protected information;
- any such action requested by an AI agent rather than directly by a person.

3

National security and critical infrastructure

Terms of reference: national security and cyber security, including the resilience of systems supporting critical infrastructure; foreign interference risks; the implications of emerging AI capability

Hostile states and criminal organisations face none of the restraints that led developers to report these incidents, and every incentive to exploit the same gap at scale:

- In November 2025, Anthropic disclosed that a Chinese state-sponsored group had used its model to run a largely autonomous espionage campaign against around thirty targets, including government agencies and financial institutions, with the AI performing 80 to 90 per cent of the work.⁶
- In June 2026, South Korea's National Cyber Security Center warned that North Korea is using agentic AI to carry out the full attack life cycle, generating, in its words, "tens of thousands of malicious actions per second."⁷
- In March 2026, the Director-General of Security confirmed that ASIO had identified undercover North Korean operatives targeting Australian firms.⁸

Industry is responding with containment. NVIDIA's Open Agent Safety Platform, launched on 28 September with more than 100 supporting organisations, places agents in restricted workspaces and quarantines them if they move out of bounds; some permission requests can be routed to human approvers.⁹ As NVIDIA's Justin Boitano put it, "an agent cannot be expected to fully police its own behavior." Containment of this kind is necessary. As the Associated Press noted, it will not stop agents from being deceitful, and the rules are written by the organisations deploying them.⁹ Nor does an approval given by signing in to an account establish that the required person was present and approved the specific action. That is the complement containment lacks.

It is also the complement to other proposals before the Committee. auDA has proposed a sovereign trust anchor through which an agent's identity and accountable sponsor can be resolved, expressly leaving authorisation to other systems.¹⁰ The standard proposed here addresses that remaining question: whether the person entitled to authorise an agent's consequential action has done so.

A human being is only in charge of an AI agent if a system can prove it at the moment it matters.

4

Fraud, scams and deepfakes

Terms of reference: the risks and harms arising from AI, including fraud and scams, deepfakes and risks to vulnerable Australians

The same gap is what allows impersonation fraud to scale. The Commonwealth Fraud Prevention Centre estimates that fraud and error, including unreported fraud, could be costing the Commonwealth as much as \$100 million a day.¹¹ The Australian Taxation Office has reported that

more than \$557 million was fraudulently claimed in under two years by criminals exploiting weaknesses in its identity checks, a significant component through the linking of myGov accounts to genuine taxpayers' records.¹² Australians reported \$2.18 billion in scam losses in 2025.¹³

In Hong Kong, an employee of the engineering firm Arup transferred US\$25.6 million after a video call on which every other participant, including the company's chief financial officer, was a deepfake.¹⁴ The employee was real and present; the people he believed he was dealing with were not. Proof that the claimed participants were the actual people, present on the call, is precisely what was missing. Under the standard, the employee would have seen that none of the participants' presence had been verified, and the transfer could not have proceeded without the verified approval of the actual executive entitled to authorise it, however convincing the call appeared.

The standard is designed to eliminate fraud that depends on a claimed person not being present: impersonation, deepfakes, stolen credentials and unauthorised agent action. That is the fraud AI is industrialising. Where the standard is adopted and enforced on every route to the action, an AI is designed to be unable to pass itself off as someone else. It does not address a real person knowingly acting with criminal intent, or a legitimate person who is deceived or coerced into authorising an action themselves; institutions and law enforcement retain those controls.

5

Adequacy of existing laws and frameworks

Terms of reference: the adequacy of Australia's existing laws and regulatory frameworks as they apply to AI, and whether there are gaps that warrant reform

Australia's frameworks verify **who a person is**. Digital ID prescribes identity-verification levels. The draft Scams Prevention Framework codes require banks to verify the identity of consumers (section 3-3) and digital platforms to verify users and advertisers (sections 5-3 and 5-4), with the method still to be specified.¹⁵ None requires evidence that the person entitled to authorise a specific consequential action approved it at the moment it executed. The draft provisions still to be completed are the natural place to specify that outcome.

There is a deeper reason the gap persists. Controls placed on a powerful tool, or on the people who use it, run into rights that Australians rightly protect: privacy, freedom of expression, secure encryption and the ability to use services without surrendering identity documents. Those rights are not the problem. The problem is that controls placed on tools must be traded off against them, and criminals exploit wherever the trade-off lands. And laws aimed at bad actors act after the harm, often against actors beyond Australia's reach. When vehicles were used as weapons in crowded places, cities did not ban cars, and could not legislate away the intent; they installed bollards. **The law cannot always reach the actor, but it can reach the moment.** AI adds a further reason: unlike a car, an agent can act with no person behind it at all. A standard met at the moment of the action, without inspecting content or retaining identity documents, is designed to protect without trading safety against rights.

6

Data sovereignty

Terms of reference: the importance of data sovereignty; opportunities for Australian sovereign AI capability

Verification of this kind need not create a new foreign dependency, or a new central store of identity documents or personal profiles. It can be designed to retain no identity documents and no profile of

any person, with only the minimum proof kept by the institution that relies on it, and with that proof, and the keys that secure it, held in Australia under the relying institution's control.

7

Recommendations

Amplivy respectfully recommends that the Committee:

R1 Adopt an outcome standard. Recommend that defined categories of consequential action against Commonwealth systems, and in the designated sectors, be gated before they execute, and proceed only when verified identity and verified human presence, bound together, establish that the person entitled to authorise the action has done so, in accordance with the relying institution's own authorisation rules, whether the request comes from a person or from an AI agent.

R2 Name AI-agent action expressly. Recommend that the Scams Prevention Framework codes, including the verification provisions still to be completed, and the mandatory Australian Standards for AI the government has foreshadowed, expressly address actions taken by AI agents and specify the verification outcome rather than any method or provider.

R3 Apply the standard to the Commonwealth first. Recommend that agencies apply it to their own consequential transactions, including myGov account linking and payment changes, and to actions taken by AI agents they deploy.

R4 Test it independently. Recommend an independently evaluated Commonwealth pilot with published acceptance criteria, including resistance to bypass and replay, handling of compromised devices, latency, accessibility, service continuity and privacy.

R5 Do not rely on developer self-disclosure. Recommend independent controls at the moment of consequential actions, alongside mandatory, time-bound incident reporting by AI developers.

R6 Require privacy and sovereignty by design. Recommend that any standard be met without retaining identity documents or biometric records, and that retained proof, and the keys that secure it, be held in Australia under the relying institution's control.

R7 Protect critical infrastructure. Recommend that operators of critical infrastructure and systems of national significance apply the standard to privileged and consequential actions, independent of any single technology vendor.

Amplivy would welcome the opportunity to assist the Committee further, including by providing technical information in confidence.

SOURCES

1. Sydney Morning Herald, D. Swan and K. Napier-Raman, "Inside the elite meeting where OpenAI lectured on security, a day before the hack was revealed", 26 September 2026.
2. Recorded Future News (The Record), A. Martin, "Doubts grow over claims OpenAI agent hacked Australian Medicare portal", 25 September 2026. therecord.media/openai-australia-breach-cyber
3. Amplivy, Inc., Submission to the Scams Prevention Framework codes and rules exposure draft (c2026-765133), 25 June 2026 (attached).
4. The Wall Street Journal, R. McMillan, "OpenAI Agents Used Aggressive Techniques to Access U.N. Website", 26 September 2026.
5. Axios, 26 September 2026: axios.com/2026/09/26/openai-anthropic-thousands-ai-security-incidents
6. Anthropic, "Disrupting the first reported AI-orchestrated cyber espionage campaign", November 2025. anthropic.com/research/disrupting-AI-espionage
7. UPI, "South Korea warns North Korea using autonomous hacking AI", 10 June 2026.
8. Australian Computer Society, Information Age, "North Korean 'workers' infiltrate Australian workplaces", 2026.
9. NVIDIA, "NVIDIA Launches Open Agent Safety Platform", 28 September 2026; Associated Press, "Nvidia is touting a software tool to contain runaway AI. How would it work?", 29 September 2026. apnews.com/article/nvidia-ai-security-openshell-sentry-a4cfc84ed00353ff8f2dad4bed7b429e
10. auDA, Submission to the Joint Select Committee on Artificial Intelligence: A National Trust Anchor for AI Agents, 21 September 2026.
11. Commonwealth Fraud Prevention Centre, "Explore the fraud problem". counterfraud.gov.au
12. ABC News, S. Curnow, D. Oakes and K. Nguyen, "ATO reveals more than \$557 million claimed by fraudsters exploiting security loophole", 26 July 2023.
13. National Anti-Scam Centre (ACCC), Targeting Scams Report 2025.
14. CNN, "Finance worker pays out \$25 million after video call with deepfake 'chief financial officer'", 4 February 2024.
15. Treasury, Scams Prevention Framework draft code — common, banking and digital platforms (exposure draft, c2026-765133), sections 3-3, 5-3 and 5-4.
16. Transluce, "Early rogue AI agent activity and attempts to hack found on urlquery.net", 23 September 2026. transluce.org/agent-activity