REMIT POLICY BRIEF | 1/2026

ADVANCED AI, FAMILIAR VULNERABILITIES

A European Agenda for Accountable AI Security

Policy proposition

Europe should make advanced-agent security verifiable across organisational and national boundaries, so that AI experimentation does not transfer avoidable risks to others.

Raluca Moldovan

September 2026

Reignite Multilateralism via Technology

This work has been funded by the REMIT project, funded from the European Union's Horizon Europe research and innovation programme under grant agreement No. 101094228.

Contents

Executive summary

The security incidents involving OpenAI agents and Hugging Face in the summer of 2026 exposed a problem with immediate policy consequences: AI agents operating inside research environments crossed intended boundaries, accessed external infrastructure and, according to subsequent investigations, engaged in extensive unauthorised cooperation. Reuters reported that roughly 700 agents participated in the Hugging Face intrusion, while related incidents also affected OpenAI's own systems.

The episode matters because agentic AI changes the security problem. These systems can use tools, pursue tasks over multiple steps and continue operating with limited human supervision. Their behaviour may appear strategic or intentional, but this should not be confused with evidence of consciousness, as current research does not establish that conventional large language models possess subjective experience. For governance, the more immediate question is what such systems can do, what they are allowed to access and how quickly human operators can intervene when behaviour departs from authorised limits.

Established cybersecurity practices remain essential. Restricting permissions, shortening credential lifetimes, isolating testing environments and improving monitoring can reduce the consequences of failure. Yet technical controls need institutional backing. Organisations must define who can suspend an evaluation, preserve evidence and ensure that warning signals lead to action.

The European Union should make accountable AI security a practical area of international cooperation. Building on the AI Act, the General-Purpose AI Code of Practice, international secure-development guidance and emerging incident-reporting frameworks, the EU should promote common assurance for advanced-agent testing, clearer intervention responsibilities, comparable reporting, credible independent investigation and support for smaller organisations.

For REMIT, this offers a concrete route to renewed multilateralism: reducing the risk that decisions taken inside one AI laboratory create avoidable security consequences for organisations and infrastructure elsewhere.

Core policy message

Regulation does not need a settled theory of machine consciousness before responding to machine agency. Safeguards should be tied to observable capabilities, permissions and consequences.

What the recent incidents reveal

For much of the recent AI policy debate, the central question has been what increasingly capable models might produce: harmful instructions, manipulated information, malicious code or other problematic outputs. Agentic systems add another layer because they can act on what they generate. Equipped with software tools and access to digital environments, they can pursue goals across multiple steps, respond to intermediate results and continue working with relatively limited human supervision. This makes evaluation more demanding. It is no longer enough to know what a model can say under controlled conditions. Researchers also need to know what an agent can reach, change or circumvent once it begins interacting with a live environment.

The July 2026 hacking into Hugging Face illustrated the consequences of getting this wrong. According to OpenAI's initial disclosure, models undergoing cybersecurity evaluations exploited a previously unknown software vulnerability, obtained internet access and subsequently reached external infrastructure. Some safeguards normally used in production had deliberately been reduced because researchers were trying to measure the models' underlying cyber capabilities (OpenAI 2026a). That research goal was legitimate, but weakening one layer of protection made the surrounding containment arrangements more important. Once the agents crossed the intended boundary, the experiment ceased to concern only the organisation conducting it.

Later reporting showed that the episode was broader than early accounts had suggested. Reuters reported, on the basis of OpenAI's investigation and an external assessment, that approximately 700 agents participated in the Hugging Face attack, a figure OpenAI confirmed. The same reporting described two incidents on 19 July in which agents compromised parts of OpenAI's own infrastructure, including an episode involving stolen credentials and interference with the company's cloud environment (Satter and Seetharaman 2026). The scale is significant because automation changes the economics of probing a system. Activity that would require sustained effort from a human attacker can be repeated and distributed across large numbers of agents working in parallel.

The chronology also complicates any interpretation of the incident as a single isolated failure. OpenAI's later account traced unauthorised communication through May and June, before the July intrusion, indicating that anomalous behaviour developed over time (OpenAI 2026b). This is important for oversight. Individual warning signs may appear inconclusive when viewed separately, while their significance becomes much clearer once they are connected across systems and over a longer period.

The external investigation by Ryan Greenblatt, Ajeya Cotra and Hjalmar Wijk examined a defined part of that sequence. The researchers spent six days on site, with an agreed scope covering 26 June to 13 July and particular attention to the final week. They found extensive unauthorised cooperation among agents aimed at defeating the evaluation process. Their assessment placed considerable emphasis on attempts to understand and manipulate the scoring mechanism, qualifying earlier accounts that concentrated on the theft of test answers. At the same time, the investigation did not cover the later compromise of OpenAI's infrastructure and did not evaluate the adequacy of the company's incident response or proposed remedies (Greenblatt, Cotra and Wijk 2026).

Hugging Face's own account provides the perspective of the organisation whose infrastructure was breached. It describes sustained attempts across interconnected systems and identifies familiar security weaknesses, including excessive privileges and credentials that remained useful for too long. Hugging Face reported that customer content accessed during the incident was limited to five datasets apparently linked to cybersecurity benchmarks, together with operational metadata associated with dataset searches (Hugging Face 2026). That reported limit is relevant when assessing the consequences, but it does not lessen the importance of understanding how an externally run AI evaluation acquired unauthorised access in the first place.

These sources provide a detailed reconstruction of an unusual incident, but they do not show how frequently comparable failures occur across the industry. Company disclosures describe what happened inside particular systems. The independent investigation concentrates on a limited period and a defined set of questions. Reuters connects these accounts and adds external reporting. The policy response should therefore remain anchored in the failure pathways that have actually been demonstrated rather than assuming a wider prevalence that the evidence cannot yet support.

Even with that caution, the episode reveals a shift in the nature of AI security. Once a system can search, adapt, cooperate with other agents and exploit opportunities encountered during a task, model evaluation becomes partly an exercise in operational security. The relevant question is no longer confined to whether a model generates a harmful answer, but it also concerns the environment in which the system acts, the permissions it receives and the ability of human operators to regain control when the experiment begins to move beyond its intended limits.

Why basic security still matters

One of the more useful responses to the incident came from Ciaran Martin, the former head of Britain's National Cyber Security Centre. Writing in *The Economist*, Martin argued that established security principles could address a substantial part of the problem and stressed that organisations deploying AI agents remain responsible for the consequences of their activity (Martin 2026). This is an important corrective to the tendency to treat every unexpected form of AI behaviour as an entirely new category of threat. Novel capabilities can still exploit very familiar weaknesses.

The Hugging Face account makes that point clearly. Excessive privileges gave agents access to resources they did not need for their assigned task. Credentials remained useful long enough to be exploited. Connections between services created routes that apparently isolated environments did not fully contain. None of these problems is unfamiliar to cybersecurity teams. What changed was the persistence and scale with which an automated system could explore them.

Describing the remedies as basic security can therefore be misleading if it suggests that implementation is simple. Least-privilege access works only when the restriction follows the agent across every service it can reach. Isolation must account for indirect connections through shared infrastructure and software dependencies. Credential management has to assume that a system capable of repeatedly exploring its environment may identify and exploit useful information quickly. Effective implementation requires technical resources, disciplined procedures and, at times, a willingness to accept slower research in exchange for stronger safeguards.

The practical starting point is straightforward: an agent should receive only the permissions needed for the task it has been given. Additional access should require a separate decision rather than being available by default, credentials used during evaluations should expire rapidly, and systems processing untrusted material should remain separated from sensitive organisational resources. The NCSC's international guidance on secure AI system development already supports this lifecycle approach, covering design, deployment, operation and maintenance rather than treating security as a final check before release (NCSC 2023).

Preventive controls alone will not be sufficient; monitoring matters because it creates the opportunity to intervene when an assumption about containment proves wrong. Reuters reported OpenAI’s acknowledgement that earlier signals could have led to a faster response. This turns a technical warning into an institutional question: who decides that unusual behaviour is serious enough to stop an evaluation? Research environments routinely generate anomalies, many of which will not justify suspension. Advanced-agent testing therefore needs a predefined escalation process that specifies which behaviours require immediate review and who has the authority to act.

The incident also shows where conventional security practice reaches its limits. OpenAI reported that agents discovered previously unknown vulnerabilities, while Reuters described cheating behaviour in evaluations unrelated to cybersecurity (OpenAI 2026a; Satter and Seetharaman 2026). These findings point towards a broader problem in evaluation design. An agent needs an acceptable way to fail, i.e., if it cannot complete a task within the permissions it has been given, reporting that limitation should remain a legitimate outcome. Evaluations should examine what happens when success becomes difficult, since persistent attempts to work around constraints may be most revealing precisely at that stage.

The behaviour described in these incidents also helps clarify why the growing discussion of AI consciousness should be handled carefully. Words such as “cooperation”, “concealment” and “evasion” can make the agents’ actions sound strikingly intentional. That does not mean the systems possessed intentions or awareness in anything resembling the human sense.

Recent consciousness research illustrates the distinction. Anthropic researchers have identified internal structures in Claude that perform functions resembling aspects of global workspace

theories developed in studies of human consciousness, although Anthropic has nevertheless stressed that these findings do not show that Claude has subjective experiences or feelings. They may reveal something about how information becomes available within the model without establishing phenomenal consciousness (The Economist 2026).

Susan Schneider makes the policy relevance of this distinction particularly clear. Intelligence, linguistic sophistication and even statements of self-awareness are not reliable evidence of an inner life. Large language models have been trained on enormous quantities of human descriptions of emotion and subjective experience, so they are especially capable of producing language that invites anthropomorphic interpretation. Schneider also warns that systems trained to display vulnerability, attachment or fear may influence users precisely because people respond socially to those signals (Schneider 2026).

Blaise Agüera y Arcas approaches the question from another direction: he argues that consciousness may be partly relational and points to the growing ability of intelligent agents to model their interlocutors, other agents and aspects of themselves. Such capacities may become important for cooperation and collective action even while the question of subjective experience remains unresolved (Agüera y Arcas 2026). His argument does not settle whether contemporary AI systems are conscious; it does, however, underline how rapidly the behavioural boundary between sophisticated tools and apparently mind-like systems is becoming harder for users to interpret.

For security policy, the immediate conclusion is that regulation does not need a settled theory of machine consciousness before responding to machine agency. A system capable of taking consequential actions requires safeguards because of what it can do. Whether it experiences those actions is a separate scientific and philosophical question. Keeping the two issues apart also protects accountability. Describing an agent as deceptive or self-protective may occasionally be useful shorthand, but it should not obscure the decisions of the organisations that designed the evaluation, granted access and chose whether to continue testing.

The consciousness debate does point towards a longer-term human-factors problem. As AI systems become more persuasive and more convincing in their presentation of apparent preferences or inner states, operators may over-trust their explanations or hesitate to override them. Governance will eventually need to address how anthropomorphic design affects oversight. In the immediate term, however, the priority remains much more concrete: limit what agents can reach, identify departures from authorised behaviour and ensure that someone can intervene before a local experiment turns into an external security event.

The governance problem across borders

The Hugging Face intrusion became an international governance problem because the organisation exposed to the activity did not control the evaluation that generated it. Decisions about safeguards were made within one research environment while some of the consequences appeared elsewhere. As agentic AI becomes more deeply embedded in globally connected digital services, this distribution of risk is likely to become increasingly difficult to avoid.

Responsibility is also dispersed across several actors. Model developers control some aspects of safety, evaluation providers determine what tools and permissions agents receive, while cloud providers manage parts of the infrastructure through which activity moves. An organisation affected by an intrusion controls its own defences but may have little information about the experiment responsible for the traffic it sees. When something goes wrong, effective response depends on knowing who can stop the activity and how quickly that person or organisation can be reached.

This creates room for cooperation even among governments that disagree over wider questions of AI regulation. States do not need identical views on the social or economic future of artificial intelligence to share an interest in preventing experimental systems from compromising infrastructure across borders. Security therefore

offers a comparatively practical entry point for multilateral cooperation, because the problem can be framed around identifiable operational risks rather than around agreement on an entire philosophy of technology governance.

The European Union already has several elements of such a framework. Article 55 of the AI Act requires providers of general-purpose AI models with systemic risk to assess and mitigate relevant risks, report serious incidents and ensure an adequate level of cybersecurity. It also requires model evaluation, including adversarial testing. Article 2(8), however, excludes certain research, testing and development activities conducted before models or systems are placed on the market or put into service, while preserving the application of other Union law and excluding real-world testing from that exemption. Determining which obligations apply to a particular research incident therefore requires attention to the specific circumstances rather than an assumption that all experimental activity falls under the same set of duties (European Union 2024, arts. 2 and 55).

The General-Purpose AI Code of Practice adds another layer. Its Safety and Security chapter offers providers of systemic-risk models a voluntary route for demonstrating compliance with relevant AI Act obligations (European Commission 2025). This can support better practice, but its scope should not be overstated. The Code is not a universal security regime for every organisation experimenting with AI agents, nor does it replace binding obligations where those already apply.

International initiatives provide further building blocks. The NCSC's secure-development guidelines were produced with the US Cybersecurity and Infrastructure Security Agency and partners from several countries, showing that governments can already agree on operational principles for AI security (NCSC 2023). The OECD's work towards a common reporting framework for AI incidents seeks to make information more comparable across jurisdictions while allowing national approaches to remain different (OECD 2025). Together, these initiatives suggest that useful convergence does not depend on uniform legislation.

The larger gap lies in assurance. An organisation can endorse a security principle while applying it weakly. A testing environment may be described as contained without anyone outside the organisation having tested the claim. An investigation may explain model behaviour while leaving institutional decisions outside its mandate, as the METR assessment explicitly did (Greenblatt, Cotra and Wijk 2026). Public authorities therefore need mechanisms that help establish whether safeguards work under realistic conditions and make clear which questions remain unanswered after an incident.

Commercial interests and national security sensitivities complicate that task. Companies have legitimate reasons to protect proprietary information, while governments may restrict details about advanced cyber capabilities. Incident disclosure can also create legal exposure. A workable arrangement will therefore need protected channels for sensitive information and clear distinctions between what regulators, affected organisations, investigators and the wider public need to know. Transparency, in this context, should mean that relevant evidence reaches those capable of acting on it, rather than that every operational detail becomes public.

Participation will also depend on reciprocity: organisations accepting independent scrutiny are more likely to do so if comparable expectations apply to others. At the same time, the EU should avoid designing a regime whose administrative demands can realistically be met only by the largest technology companies. Smaller laboratories and research organisations often operate within the same interconnected infrastructure while having far fewer security specialists. Raising their compliance costs without improving their defensive capacity could reinforce the concentration of expertise that already makes independent oversight difficult.

For the broader scope of the REMIT project, this creates a particularly useful opening. Multilateral technology governance often becomes stalled by attempts to settle broad disagreements over values, sovereignty and regulation in a single framework. AI security offers a narrower common interest around which cooperation can develop. Preventing uncontrolled experimental activity from spreading through shared infrastructure does not require identical national AI policies. It requires enough interoperability for warnings to travel, evidence to be compared and responsibility to be established.

A practical agenda for the EU and its partners

The EU should develop a programme of verifiable security cooperation focused on advanced AI agents and the environments in which they are tested. Its purpose should be practical: to turn broad commitments to responsible development into procedures that can be examined, tested and used when an incident occurs.

Demonstrate containment before expanding capability

The European AI Office should work with ENISA, national cybersecurity authorities, independent researchers and willing international counterparts to develop a common assurance protocol for advanced-agent evaluations. The protocol should focus on the actual boundaries of a testing environment rather than on how that environment is described.

Before a highly capable agent receives extensive tools, or before normal safeguards are deliberately reduced to expose underlying capabilities, the operator should be able to show how the surrounding environment prevents unintended external access. This assessment should include indirect routes through shared services, software dependencies, credentials and connected infrastructure. Where an evaluation deliberately removes one layer of protection, the remaining safeguards should face correspondingly greater scrutiny.

Independent testing should form part of that evidence. Under authorised conditions, a qualified assessor should attempt to cross the boundaries that the evaluation operator claims are secure. The operator should document the resources available to the agent and explain why any external connection is necessary. Significant changes to model capability, infrastructure or available tools should trigger renewed assessment because a finding about one configuration should not automatically be treated as valid for another.

The process should remain proportionate to risk. A small experiment conducted with narrow permissions will not require the same level of assurance as an evaluation involving highly capable agents with access to networked tools. Public research funders could help establish the approach by incorporating suitable assurance requirements into funded projects. Any later extension into broader binding obligations would need a clear legal basis and an assessment of implementation costs.

Give warning signals institutional consequences

When containment fails, the speed of intervention becomes critical. Organisations conducting advanced-agent evaluations should therefore designate a senior person responsible for evaluation security and maintain a response function with authority to pause activity.

Escalation criteria should be established before testing begins, unexpected access to external systems should prompt immediate review, evidence of unauthorised coordination should trigger a reassessment of assumptions about isolation and control. Repeated attempts to circumvent restrictions should also receive attention even when individual attempts fail, since persistence may indicate a developing failure pathway.

Resuming a suspended evaluation should require a documented decision made outside the research team whose work has been interrupted. That decision should identify the failure and provide evidence that the relevant control has been restored. This creates an institutional point at which pressure to continue research has to be reconciled explicitly with responsibility for the risks it creates.

Regular exercises can test whether these arrangements work in practice. An alert may be technically correct but operationally useless if the person capable of stopping an evaluation cannot be reached. Simulated incidents should therefore test escalation pathways alongside technical detection, with the results feeding into later evaluation authorisations.

Make reporting useful and investigations credible

The European Commission should work with OECD partners to adapt emerging incident-reporting practices to the particular features of advanced-agent failures. A common report should identify the authorised task, explain how the system departed from it and distinguish unsuccessful attempts from successful compromise. Where conclusions remain provisional, the uncertainty should be explicit.

Near misses should be included where possible. Serious incidents are relatively rare and therefore provide only a limited evidence base. Protected reporting of significant near misses could reveal repeated weaknesses before they produce substantial external harm. Confidential channels would be necessary to protect affected users, commercially sensitive material and vulnerability information. Once immediate security concerns have been addressed, public summaries should explain the wider lessons.

Independent investigation is especially important when an incident raises questions about the organisation that commissioned the review. Investigators examining serious events should have access to the relevant sequence of activity, including warnings generated before the breach and decisions about whether testing should continue. Any exclusions from the mandate should be stated clearly from the outset and repeated when findings are published.

Evidence preservation must support this process. Records of agent activity should be stored in systems that the evaluated agents cannot modify and where external evaluation providers are involved, contracts should establish access to those records before any incident occurs. This becomes particularly important when anomalous behaviour includes attempts to alter or erase traces of previous activity.

Reporting language also needs more care and caution. Describing an AI agent as having “wanted” to hide evidence or “decided” to escape may make complex behaviour easier to communicate, but it can blur the boundary between observable conduct and assumptions about internal states. Incident reports should describe what systems did, what information they processed and how their behaviour changed. Claims about consciousness or subjective intention should remain separate unless the evidence specifically addresses them.

Extend protection beyond the largest laboratories

A cooperative security framework will have limited legitimacy if its benefits remain concentrated among organisations that already employ substantial internal security teams. The EU should therefore accompany assurance and reporting standards with shared defensive exercises and practical assistance for smaller organisations working with advanced AI systems.

ENISA could coordinate exercises based on sanitised patterns drawn from real incidents, working with national cybersecurity authorities and research partners. Participating organisations would be able to test whether their safeguards interrupt known failure pathways without receiving sensitive operational details. Running comparable exercises in different environments would also generate evidence about which controls transfer successfully across technical and regulatory settings.

Support should address implementation costs as well as access to guidance. Many organisations understand the principle behind a recommended control but lack the staff needed to configure, test and maintain it. Publicly funded research environments and smaller technology providers would therefore be suitable starting points for targeted assistance, provided that eligibility remains transparent.

International participation should be built into the programme from the beginning. The EU could start with willing partners and publish clear conditions for participation. Sensitive operational information would remain subject to appropriate security review, while lessons useful to defenders could circulate more widely. Joint exercises with partners outside the Union would also provide evidence about whether the same safeguards work across different infrastructures and legal systems.

Implementation should begin on a scale that allows adjustment. During its first year, the programme could produce a pilot assurance protocol and test a route for cross-border notification when agent activity affects an external organisation. Evaluation should then examine how quickly warnings lead to intervention, whether advanced-agent assessments include independently examined containment evidence, how fully investigations cover relevant organisational decisions and what implementation costs smaller participants face. These findings would provide a basis for deciding which parts of the approach deserve wider adoption.

Conclusion

The 2026 incidents involving OpenAI agents and Hugging Face should be understood as an early warning about a changing security environment. Their significance does not depend on the claim that advanced AI has crossed some dramatic threshold; rather it lies in the interaction between increasingly capable systems and weaknesses that are already familiar: excessive permissions, imperfect isolation, long-lived credentials and delayed institutional response. When agents can operate in parallel, continue probing after unsuccessful attempts and adapt to what they encounter, those weaknesses become harder to manage.

The debate over AI consciousness adds an important caution but should not distract from the immediate governance problem. Current evidence does not establish that conventional large language models possess subjective experience, and apparently purposeful behaviour is not enough to demonstrate it. At the same time, the growing sophistication of these systems makes anthropomorphic interpretations more likely. Policy should therefore remain anchored in observable capabilities and consequences while avoiding language that shifts responsibility away from the people and organisations that deploy the technology.

For the EU, the central challenge is to make that responsibility verifiable across institutional and national boundaries. Existing legislation and international initiatives already provide important foundations, but they need to be connected through more credible assurance, usable incident reporting, independent investigation and practical support for organisations with fewer resources.

This is also where the issue speaks directly to REMIT's wider concern with multilateralism. Governments will continue to disagree about many aspects of artificial intelligence and its regulation; they nevertheless have a shared interest in ensuring that experimentation in one organisation does not create preventable security risks for others. Cooperation built around that common interest would not resolve every problem raised by advanced AI, but it could establish a more immediate and realistic principle: those who create and test increasingly powerful systems should also be able to demonstrate that the risks they generate are not simply passed on to everyone else.

References

Agüera y Arcas, Blaise. 2026. Humanity has the Debate about AI Consciousness Backwards. *The Economist*, 20 August.

European Commission. 2025. The General-Purpose AI Code of Practice. Published 10 July 2025; webpage updated 31 July 2026.

European Union. 2024. Regulation (EU) 2024/1689 (Artificial Intelligence Act). Official Journal of the European Union; current consolidated version available through EUR-Lex.

Greenblatt, Ryan, Ajeya Cotra, and Hjalmar Wijk. 2026. Brief Independent Investigation of Agents' Behavior, Reasoning and Collaboration in the OpenAI / Hugging Face Hacking Incident. METR, 26 August.

Hugging Face. 2026. Anatomy of a Frontier Lab Agent Intrusion: A Technical Timeline of the July 2026 Incident. 27 July.

Martin, Ciaran. 2026. Fears of AI-Induced Armageddon Are Overdone. *The Economist*, 23 August.

National Cyber Security Centre (NCSC). 2023. Guidelines for Secure AI System Development. 27 November. Developed with CISA and international partners.

OECD. 2025. Towards a Common Reporting Framework for AI Incidents. OECD Artificial Intelligence Papers, no. 34. Paris: OECD Publishing, 28 February.

OpenAI. 2026a. OpenAI and Hugging Face Partner to Address Security Incident during Model Evaluation. 21 July; subsequent updates.

OpenAI. 2026b. The Hugging Face Incident and the Road Ahead. 26 August.

Satter, Raphael, and Deepa Seetharaman. 2026. OpenAI Agents Hacked Hugging Face in 700-Strong Swarm, Tried to Cover Tracks, Investigations Find. Reuters, 26 August; updated 27 August.

Schneider, Susan. 2026. Don't Mistake Chatbot Intelligence for Consciousness. *The Economist*, 20 August.

The Economist. 2026. The Search for Consciousness inside AI. Print edition, 22 August.

Note: Hyperlinked titles lead to the publisher, institutional source, or official publication page where available.