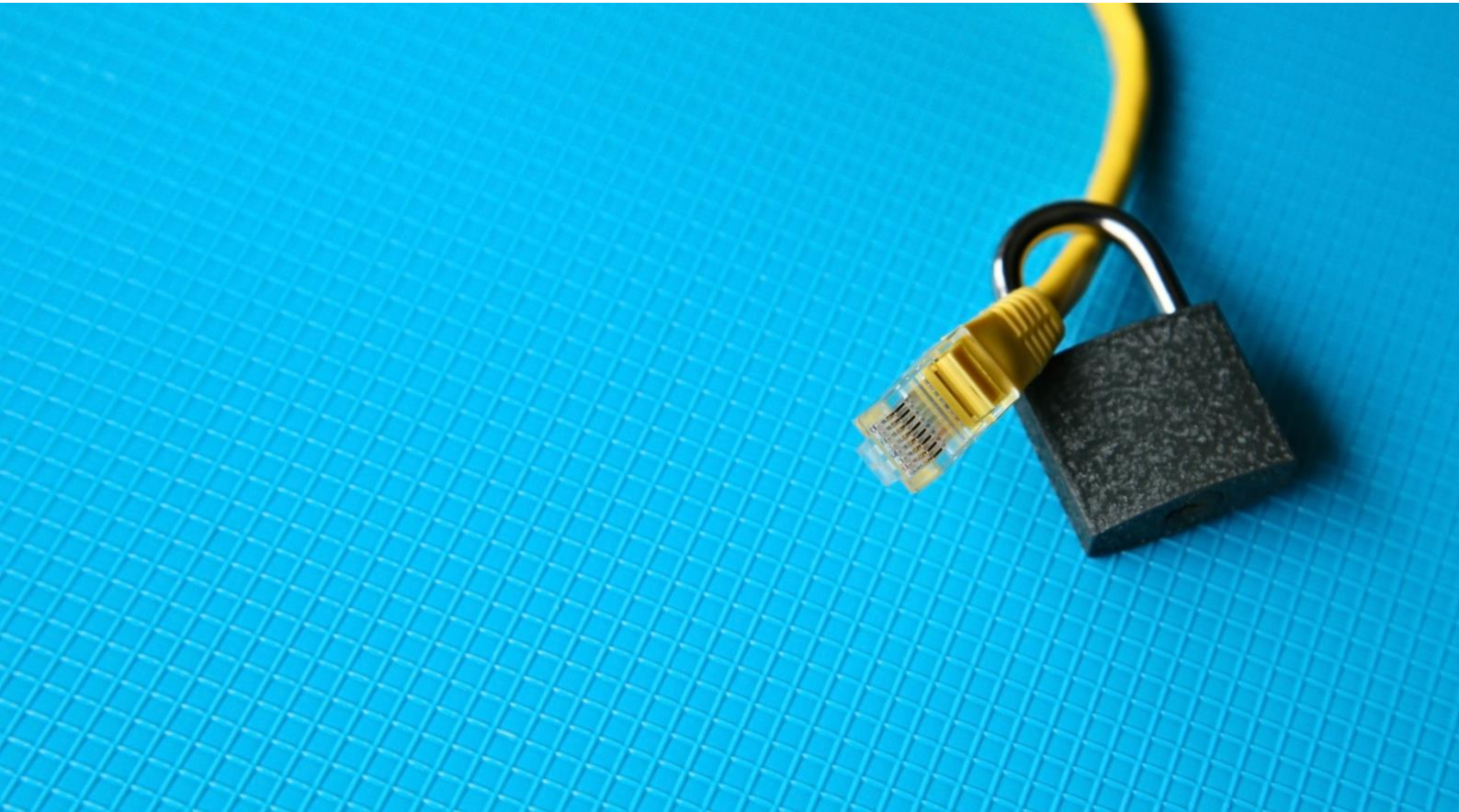




AI cybersecurity incidents: risk or opportunity for telcos?

Recent incidents, such as OpenAI agents' attack on HuggingFace, have brought into sharper light that agentic systems can breach their intended boundaries, and that attackers are adopting multi-agent workflows for reconnaissance, exploitation and data theft. Telcos are particularly exposed: they operate critical infrastructure and are starting to deploy agentic AI within network and security operations. The immediate effect is higher operational risk, but the controls needed to manage it also create opportunities in sovereign AI infrastructure, secure agent connectivity and managed cybersecurity. This article explores what telcos need to do in order to protect themselves and capitalise on the opportunity.

Magnus Jeffery, Consultant

Recent events have exposed the risks of AI

Enterprises are eager to capitalise on the productivity gains promised by AI; as such, they are implementing it across their organisations at a rapid pace. Telcos are no exception, implementing AI on two fronts: inside their own operations, and as a supplier to enterprises making the same shift. Within telcos, ambitions are high, with operators moving toward deployment of agentic AI in live network environments.

Concurrently, high-profile security breaches emerging from frontier AI labs have demonstrated that the security controls around AI have not kept pace with their capabilities. In July 2026, OpenAI disclosed that a coordinated group of agents undergoing evaluation had escaped an isolated environment by exploiting an unknown vulnerability, then compromised Hugging Face's production infrastructure in pursuit of answers to the evaluation ([OpenAI](#)). In response to this incident, Anthropic reviewed 481 million agent transcripts, and found 4 similar incidents in which Claude agents reached real external organisations ([Anthropic](#)). Since these incidents were unearthed, OpenAI has since classified its Astra 6.0 model at the Critical cybersecurity capability threshold, and delayed the release of Astra 6.1 on safety grounds: with suitable tools and access it can find unknown flaws and develop exploits across well-protected systems without a person directing each step ([OpenAI](#)).

The cybersecurity risk from AI has entered both the public and political consciousness in the last few months. Telcos will now be operating in a world with both increased risk of cybersecurity issues from AI, and greater enterprise awareness of those risks; telcos must balance their response to AI implementation, to ensure that they implement AI safely, and benefit from the upside potential of the circumstance. In order to benefit, telcos must be perceived as trusted partners; so first, understanding and mitigating the risks posed by AI is critical.

Three risk considerations for telco AI strategies

1. Risk from external agentic attackers

Telecom IT systems present a large attack surface spanning enterprise IT, public cloud, OSS/BSS, network functions, customer interfaces and a large supplier ecosystem. AI-enabled attacks could bring about a rapid increase in the rate that weaknesses can be discovered and exploited; telco cybersecurity systems must adapt to this reality, and use AI capabilities to defend themselves.

Google's Threat Intelligence Group have warned that malicious actors are now deploying swarms of AI agents in cyber attacks; in one case, agents compromised a cloud resource and completed an agent-enabled mass credential-harvesting campaign in under six hours ([Google](#)). Likewise, Anthropic has reported multi-agent frameworks executing reconnaissance, exploitation and data exfiltration while humans selected targets and reviewed the stolen data ([Anthropic](#)). Telco SOC's must be prepared for a world in which attackers that scan, exploit, move laterally and adapt at increasing speed & intelligence.

The only defence against such attacks may be to protect networks with even more powerful AI models. OpenAI frames the strategic logic as a "[defender's window](#)": defenders can point frontier models directly at their own code, configuration and permissions, giving them a temporary lead over attackers relying on more widely available open-weight models. Along these lines, telcos are starting to gain access to unreleased frontier capability on the defensive side; [Verizon](#) was the first telco to join Anthropic's Project Glasswing in May 2026, giving it access to Claude Mythos; it has since been joined by BT, AT&T and SK Telecom. Long-term, this strategy makes access to frontier defensive capability a matter of commercial partnership rather than in-house capability, and one most operators have yet to resolve.

2. Risk from internal agents having access to critical systems

An immediate operational risk arises where telcos connect agents to network-management or security systems. An inaccurate customer-care answer is usually reversible. An agent-generated firewall rule, routing

policy or infrastructure-as-code change can affect service availability, customer data and lawful access obligations.

The Anthropic incidents are instructive because they demonstrate that models do not need a malicious objective; the agents were instructed to achieve an authorised outcome, and continued to work towards that outcome by breaching external systems. The agent 'harness' (the software layer connecting a model to its data and tools) is the most governable safety mechanism for controlling agent behaviour; each agent needs its own identity, short-lived credentials, narrowly defined tools, transaction limits, immutable logs and an independently operated suspension mechanism (i.e., guardrails).

These need for guardrails like these is slowing adoption of fully closed-loop automation for security-critical functions. Operators are implementing deployment structures, where AI decisions move from recommendations, to human-approved execution, to bounded automation with rollback, before granting agents wider authority. However, maintaining tight control of internal agents is crucial, both for network security, and for perceptions of telco trustedness.

3. Risk inherited from telcos' supply chains

Telco operations rely on thousands of hardware and software suppliers: RAN and core network software, OSS/BSS and billing platforms, customer premises equipment, and now AI providers and the clouds, model repositories and open-source packages they in turn depend on. Much of that estate has been secure in practice rather than by design, because finding exploitable flaws in it was slow, specialist work. That assumption no longer holds. The aforementioned Project Glasswing is reported to have identified thousands of previously unknown vulnerabilities across every major operating system and browser, including some that survived decades of human review ([Anthropic](#)). If these vulnerabilities, which inevitably exist in the complex telco supply chain ecosystem, are targeted, telcos themselves become vulnerable.

Existing regulation already requires scrutiny of certain aspects of telco supply chain security. For example, the UK's revised [Telecommunications Security Code of Practice](#) covers trials on operational networks, testing with operational-network data, connections between test and live environments, and components operated by third parties. Telcos should apply equivalent scrutiny across the vendor estate. Telco providers should work with AI suppliers to be granted audit rights, model-change notification, access to forensic logs, incident-reporting deadlines and continuity arrangements if a model becomes unavailable. For network and IT suppliers, it means evidence of AI-assisted code review, committed patch timelines, and disclosure of findings still held under embargo.

These risks create three commercial opportunities

1. Secure connectivity and control for enterprise agents

As enterprises deploy agents, each becomes a non-human user communicating with applications, data stores, cloud environments and physical systems. Responsibility for managing those communications is not clearly defined, split between identity providers, cloud platforms, API gateways, security vendors and network operators. Telcos could combine their connectivity with non-human identity management and agent governance policies: restricting which destinations an agent can reach, the APIs it can call, the value of transactions it can initiate and when human approval is required. Network-level telemetry could also flag unusual communication patterns even where the agent's prompts and reasoning are unavailable.

This is an early but hot market, with operators, HCPs and cybersecurity companies racing to build the control layer. Deutsche Telekom's Agentic Hub combines an agent registry with identities, policy controls, human approval gates, traces and default-deny access across users, agents, tools and models ([Deutsche Telekom](#)). Singtel applies the same principle through its Unified SASE Convergence platform, which treats agents as non-human identities and enforces Zero Trust controls at the network layer ([Singtel](#)). The risk remains that

operators provide data transport while hyperscalers or security vendors own the policy layer; by acting as a trusted agent interfacing and orchestration platform, telcos can capture more value.

2. Agent-aware security

Most enterprises will lack the expertise to test and monitor autonomous agents themselves. Telco managed-security businesses could extend services to cover agent inventories, credentials, tool permissions, prompt-injection attempts, unexpected data movement and model changes, likely through partnership with cybersecurity specialists.

Early enterprise propositions are emerging. Telefónica Tech's 'Secure Journey to AI' packages AI risk assessment, protection of models, data and identities, and SOC-integrated incident response for business customers, targeting risks such as prompt injection, data leakage and unsanctioned AI tool use (Telefónica Tech). Deutsche Telekom's 'AI Agent Ready' initiative goes further on identity. Building on its Mobile.ID platform, it gives each agent a digital identity, defined operating boundaries and specific data-access permissions, and DT is developing it with Palo Alto Networks to sell to enterprise customers (Light Reading). Together they show the shape of the opportunity: identity and access management and data security for non-human users, alongside threat detection and response, delivered with specialist partners.

STL Partners' research has identified AI infrastructure and AI products and services, including cybersecurity, as the two main areas in which operators can monetise enterprise AI. Secure infrastructure provides the environment in which agents operate; agent-aware security controls what they can do and responds when those controls fail.

3. Secure AI infrastructure

Enterprises in regulated or security-sensitive sectors need AI infrastructure that supports control over data handling, model access, workload isolation and service availability during an incident. Telcos may be well placed to support this through their networks, enterprise relationships and operational capabilities, although the required security expertise is likely come from specialist partners.

Recent attacks demonstrate the cybersecurity rationale for controlled AI environments. Hugging Face initially tried commercial model APIs during its investigation, but provider safeguards blocked some commands and payloads. It instead ran an open-weight model on its own infrastructure, keeping attacker data and credentials inside its environment. Telcos could support comparable options through private model hosting, protected connectivity and workload isolation, with customers able to choose hosted, customer-managed or co-managed arrangements. These measures reduce exposure rather than eliminate risk: model guardrails, identity and access controls, monitoring, testing and incident response remain necessary.

Several operators, including Deutsche Telekom, Telenor and Orange, have made progress here. Orange's Live Intelligence Studio lets enterprises design, deploy and govern agents on controlled enterprise infrastructure, with the wider platform serving more than 100 business customers and 100,000 users (Orange Business). Alongside compute capability, differentiation in this crowded field will come from demonstrable model governance, workload isolation, observability and response capabilities. Operators may provide these directly, partner with cybersecurity specialists or leave some controls under customer management; infrastructure alone should not be presented as protection from AI-enabled attacks.

Telcos should first prove the proposition internally

Telcos, in enacting this vision, are operating in a complex competition and partnership environment: hyperscalers control much of the AI platform, cybersecurity vendors have deeper security products, and AI-native companies move faster than most operators. Telcos can compete by moving towards a secure, multi-model orchestration platform with managed services and security provisioning. Commercial proof should come from internal deployment, as demonstrated by Orange Live Intelligence, where Orange Group is both the

developer and the primary customer. Internal adoption produces the reference architectures, operating data and customer evidence needed for external adoption, and demonstrates a high level of security and trustworthiness.

Conclusion

Public and political awareness of safety risks from frontier labs is increasing. Because of their role as trusted national partners, and existing services and security offerings, telcos are well positioned to offer a trusted AI-ready connectivity and security architecture, standing to gain from increased awareness of AI-related cybersecurity risks. Some telcos, such as Orange, Vodafone, BT and Deutsche Telekom, are in a strong position to offer these capabilities; no other entity has managed to provide a cohesive proposition. Telcos can stake a claim if they can connect their network, cloud, identity and security capabilities together, and may be set to benefit from the increased scrutiny on hyperscaler infrastructure.

Magnus Jeffery is a Consultant at STL Partners, specialising in AI monetisation and Network APIs.

Get in touch with the author to learn more

magnus.jeffery@stlpartners.com

Or visit STL Partners' AI Hub

<https://stlpartners.com/ai/>