Hogan Lovells

Insights and Analysis

International AI Safety Report 2026 – What it means for business crime risk

General-purpose AI is making fraud, impersonation and cyber compromise cheaper, faster and harder to attribute – and current safeguards do not reliably prevent harm.

18 February 2026

The International AI Safety Report 2026 is framed as a scientific assessment to support policymaking. It is organised around three questions: what general-purpose AI can do today, what emerging risks it poses, and what risk management approaches exist – and how effective they are.

For corporates, the immediate issue is misuse: AI-assisted content and operations that erode trust-based controls. The report also notes the scale and unevenness of adoption – around 700 million weekly users of leading systems, with some countries above 50% usage, while parts of Africa, Asia and Latin America likely remain below 10% – meaning risk maturity will vary across markets and third-party ecosystems.

Read through a business crime lens, the report points to three practical problems: it is easier than ever to defraud a corporate; it is harder than ever to check whether something is real; and it is unclear who you can hold accountable when the harm occurs. Governance initiatives are proliferating (including EU work on general-purpose AI, the G7 Hiroshima AI Process reporting framework, and safety frameworks published by major developers), but "evidence on real-world effectiveness of most risk management measures remains limited".

Analysis

1. Synthetic content is a direct threat to fraud controls

The report highlights harmful incidents involving AI-generated content, particularly audio and video impersonation. It records research suggesting listeners can mistake AI-generated voices for real speakers 80% of the time, and it describes cases where cloned voices were used to persuade victims to transfer money by exploiting trust-based approval processes.

This is a real and immediate threat to businesses: a credible "executive" request to move funds, change supplier bank details, override steps, reset credentials, or share sensitive information. The risk is amplified by three features: low cost and low skill requirements; difficulty verifying authenticity under time pressure; and imperfect detection.

The report notes the limits of technical fixes – "watermarks and labels can help ⋯ but skilled actors can often remove them" – and that "identifying where deepfakes come from is also difficult".

2. Cyber operations are being commoditised – even without full autonomy

The report devotes substantial attention to AI use in cyberattacks, noting that developers increasingly report misuse of their systems in cyber operations and that illicit marketplaces sell easy-to-use tools that lower attacker skill requirements.

It is careful about overclaiming. Fully autonomous end-to-end cyberattacks have not been reported, and it is difficult to confirm whether real-world incident levels have increased because of AI. The

more practical point is that blended attacks are becoming easier: synthetic content to gain initial access or authorisation, paired with AI-assisted exploitation and persistence.

The report also offers a measured note of optimism: it remains an open question "whether future capability improvements will benefit attackers or defenders more". But that advantage will only materialise where organisations deploy AI effectively in security and fraud detection.

3. Malfunction, misstatement and automation bias create exposure

The business crime story here is not only external victimisation. General-purpose AI can also enable fraud from within – employees, agents and other "associated persons" generating credible fakes and muddying audit trails.

In the UK, that intersects with the "failure to prevent fraud" offence (in force from 1 September 2025 for large organisations), which focuses on whether an organisation had reasonable prevention procedures when an associated person commits specified fraud offences intending to benefit the organisation (or its clients). AI-enabled methods – including fake approval requests, fabricated supporting documents and high-volume communications – should now be considered in any fraud risk assessment, training and controls.

The report's discussion of reliability challenges and "automation bias" adds a separate exposure route: where teams defer to AI-assisted outputs "even when they are wrong", organisations risk making decisions and statements that are harder to evidence and defend – including in disclosures, contractual representations and dealings with regulators or counterparties.

4. Risk management is improving, but remediation remains uncertain

Over half of the report is dedicated to risk management practices. It observes that "most risk management initiatives remain voluntary, but a few jurisdictions are beginning to formalise some practices as legal requirements".

The report's assessment is cautious: current measures do not reliably prevent harm, and evidence of

effectiveness in real-world conditions remains limited. It also highlights uncertainty over liability allocation because harms can be difficult to trace to specific design choices and responsibilities are distributed across multiple actors.

The report therefore stresses "defence-in-depth" – multiple layers of safeguards "so that if one safeguard fails, other safeguards may still prevent harm" – and "building societal resilience", defined as "the ability of societal systems to resist, absorb, recover from and adapt to shocks and harms". For businesses, the lesson is simple: rehearse AI incident response in practical detail – who leads, who sits on the incident response team, who has authority to pause or shut down affected systems if needed, and how you will communicate externally about the incident.

Practical takeaways

- Treat voice, video and email as untrusted for high-risk actions (payments, supplier changes, credential resets, urgent approvals) and require out-of-band verification.
- Build "defence in depth": independent safeguards so that if one fails, another still prevents harm (dual approvals, call-backs via known numbers, friction for first-time or changed payees).
- Run AI incident scenarios: contested authenticity, rapid shut-down decisions, and evidence preservation.
- Update fraud risk assessments and "reasonable procedures" programmes to cover AI-enabled methods (synthetic media, document fabrication, insider misuse), and extend that approach to key third parties. This is required under the new failure to prevent fraud offence.
- Pressure-test accountability and recourse in contracts and governance with AI vendors and critical suppliers, assuming remediation may be slow or impossible.

Conclusion

The report is aimed at policymakers, but its corporate message is immediate: general-purpose AI means the barrier to entry for fraud and deception is lower than ever before.

The sensible response is resilience: building “defence in depth”, treating AI-assisted fraud as part of your prevention programme and rehearsing how your business will respond to a critical incident.

Please get in touch if you would like support with incident response planning or crisis simulation. We have developed an immersive training exercise that drops leadership teams into a fast-moving AI incident – with unfolding facts, audit and compliance pressure – so you can see where your controls hold or break.

Authored by Reuben Vandercruyssen, Liam Naidoo and Alex Cumming.

Contacts

Reuben Vandercruyssen

Senior Associate

London

Email me

Liam Naidoo

Partner

London

Email me