

PROTECTION AGAINST AI ASSISTED THREATS

Review verification, staff awareness, and technical defenses as outside threats change.

Client Education Series

An exclusive publication of Sandbox Technologies, Inc.

Sandbox Technologies
3868 W Carson St, Suite 316
Torrance, CA 90503
www.sandboxtech.com
Tel. 424.207.5130

SANDBOX TECHNOLOGIES

Small Business Solutions
Enterprise GrowthPath®
EGP Secure®
EGP Cloud™
ConstructIT®

Protection against AI Assisted Threats

Review verification, staff awareness, and technical defenses as outside threats change.

Review the decisions an attacker could influence

An apparently convincing message or familiar voice should not replace an established verification procedure. Review requests involving payments, account changes, credentials, sensitive information, or urgent exceptions. Confirm unusual requests through a known contact method rather than details supplied in the request itself.

Staff should know how to pause a transaction, report a concern, and obtain a decision without being pressured to proceed. Practice with examples that reflect the requests employees actually receive.

Maintain the underlying protections

Revisit account security, email filtering, device protection, access permissions, software updates, and recovery arrangements. AI-related risks should be considered within the broader security program. Confirm that the controls you rely on are in place and working, including for accounts and applications introduced during AI adoption.

Where AI can read outside messages, documents, or websites, consider the risk of malicious content steering its behavior. Limit what the system can access and change, keep retrieved material separate from authority to act, and require approval for consequential actions. An AI summary should not be treated as independent verification of the source.

Test procedures and prepare to respond

Review awareness training and verification exercises routinely, even if earlier guidance has been followed. Check that employees know who to contact and that requests are handled consistently when key staff are unavailable.

If a suspected incident occurs, follow the organization's response process. Preserve relevant records, contain access as appropriate, and involve the responsible security staff. Review what failed before restoring or expanding affected workflows. The NCSC guidance linked below discusses AI system risks, including incorrect output and prompt injection, and the need for security throughout the system's life.

We're here to help and have structured services to address each of these areas individually or as part of a coordinated plan. To discuss what deserves attention first, contact your Sandbox Technologies Account Manager or Consulting CIO.

Further reading: NCSC, AI and cyber security: what you need to know

https://www.ncsc.gov.uk/guidance/ai-and-cyber-security-what-you-need-to-know

SANDBOX TECHNOLOGIES

Copyright 2026, Sandbox Technologies, Inc.