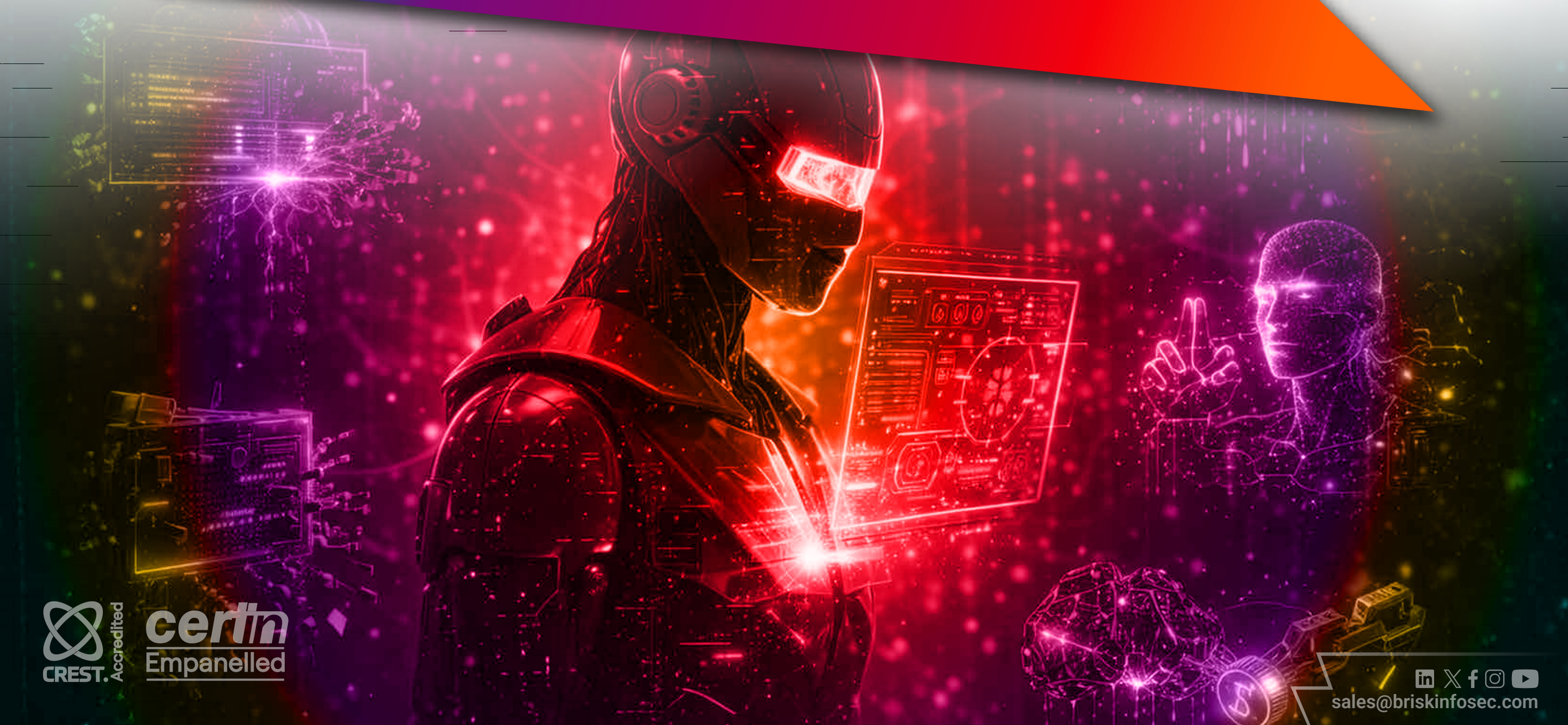
6 AI THREATS TO KNOW IN 2026

- AI is no longer just a tool for productivity. It is now embedded in attack automation, deception, and exploitation.
- This deck outlines six AI-driven threats that are changing how intrusions scale, evade, and succeed.



CREST Accredited

certin Empanelled

sales@briskinfosec.com

AGENTIC AI & “SHADOW AGENT” INTRUSIONS

Autonomous AI agents capable of executing multi-stage cyber operations with minimal human oversight. Internally, unsanctioned “Shadow Agents” connected by employees create unmonitored access paths to enterprise data.

About

Adversaries weaponize agentic systems to scan, exploit, and exfiltrate autonomously

AI executes reconnaissance → exploitation → data extraction without pause

Employees integrate AI agents into workflows without governance, exposing internal APIs and data stores

How It Works

Risk

80–90% of intrusion lifecycle executed without manual control

Attack speed exceeds human response cycles

Shadow AI increases detection time and breach cost

Invisible attack surfaces outside formal security controls

CREST Accredited

certin Empanelled

sales@briskinfosec.com

PROMPT INJECTION & ZERO-CLICK EXPLOITS

Malicious instructions embedded within AI-readable content that manipulate the model into bypassing guardrails. Evolved into “zero-click” attacks requiring no user action.

About

Invisible instructions embedded in emails, documents, or web content

AI assistant processes the content automatically

The model is coerced into executing unintended actions

Example: EchoLeak (CVE-2025-32711) enabling automated data exfiltration

How It Works

Risk

AI assistants act as “confused deputies”

Data exfiltration from connected systems (OneDrive, Teams, internal databases)

Security controls bypassed without phishing clicks

Traditional email security becomes insufficient

CREST Accredited

certin Empanelled

sales@briskinfosec.com

ADAPTIVE & SELF-MUTATING MALWARE

Malware using embedded AI models to detect its environment and rewrite its code dynamically.

About

How It Works

Risk

- Generates scripts in real time using local LLMs
- Detects sandbox or analysis environments
- Alters payload behavior to evade detection
- Can “play dead” during forensic inspection

- Signature-based defenses fail
- Behavioral detection models lose reliability
- Extended dwell time
- Increased incident response complexity

CREST Accredited

certin Empanelled

sales@briskinfosec.com

HYPER-PERSONALIZED SOCIAL ENGINEERING ("VIBE HACKING")

AI-driven cloning of voice, writing style, and digital persona to create highly convincing targeted deception.

About

Scrapes public and internal signals

Replicates tone, vocabulary, and communication patterns

Voice cloning possible from 3–5 seconds of audio

Deepfake video conferencing used in high-value fraud cases

How It Works

Risk

Financial fraud at executive level

Vendor payment redirection

Breakdown of voice/video trust

Identity verification frameworks become obsolete

CREST Accredited

certin Empanelled

sales@briskinfosec.com

SUPPLY CHAIN & OAUTH TOKEN ABUSE

Attackers compromise trusted SaaS integrations and non-human identities instead of breaching systems directly.

About

How It Works

Risk

- Theft of OAuth tokens
- Abuse of API keys and service accounts
- Access granted as legitimate third-party applications
- Example: Drift/Salesforce breach impacting 700+ organizations
- Perimeter bypass using valid credentials
- Excessive permissions on service identities
- Lateral movement across SaaS ecosystems
- Low visibility into machine-to-machine authentication

CREST Accredited

certin Empanelled

sales@briskinfosec.com

DATA & MODEL POISONING

Compromising AI integrity by injecting malicious data into training sets or retrieval-augmented generation (RAG) sources.

About

How It Works

- Insertion of malicious instructions into knowledge bases
- Sleeper triggers activated under specific prompts
- Examples: PoisonGPT, Morris II worm replication via poisoned RAG

Risk

- AI outputs manipulated without obvious failure
- Sensitive data leakage on trigger
- Strategic decision distortion
- Erosion of trust in AI-assisted workflows

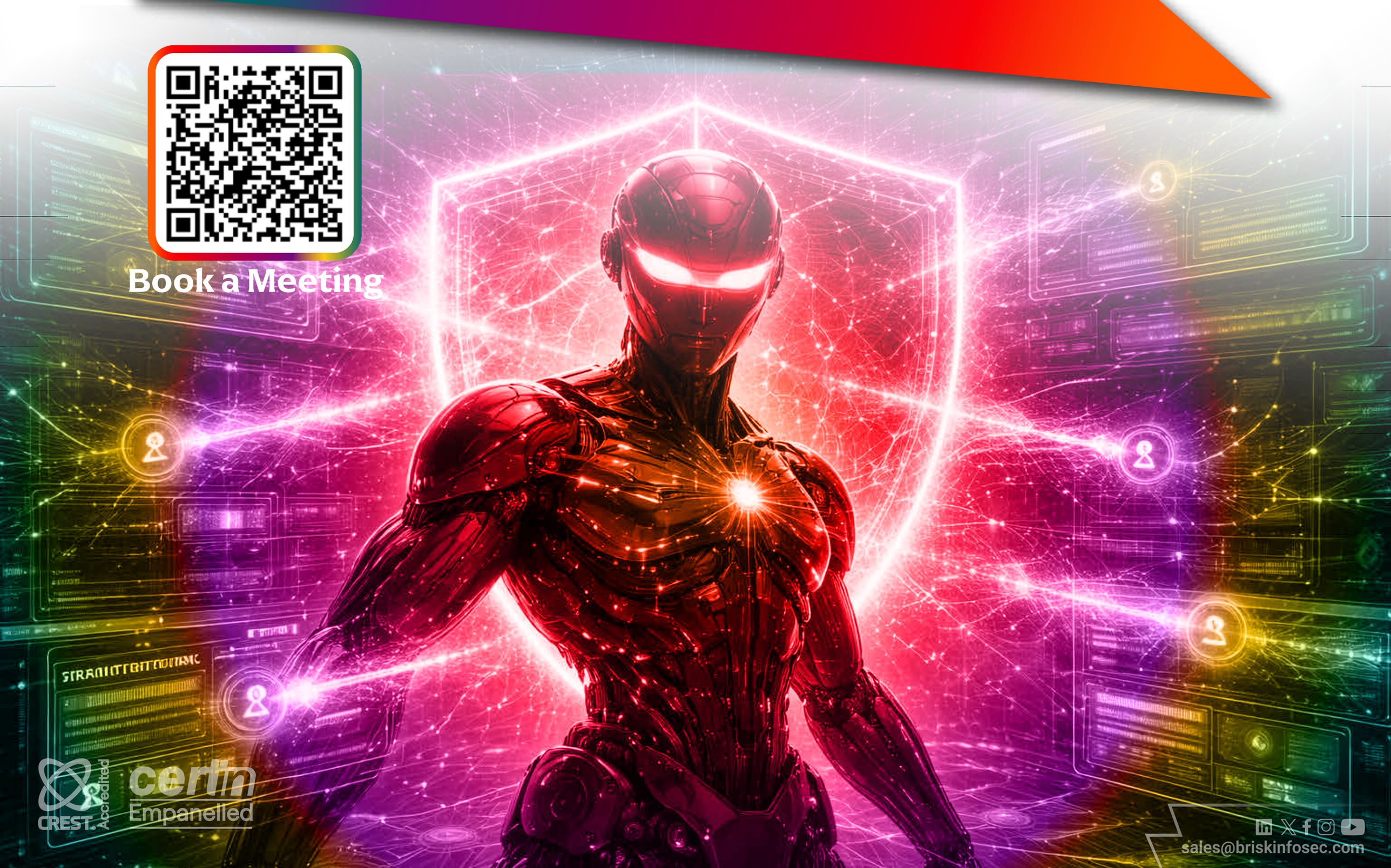
sales@briskinfosec.com

BriskInfosec™

TEST YOUR READINESS AGAINST AI-POWERED ATTACKS

- If attackers can automate intrusion cycles,your defenses must be validated the same way.
- Simulate. Identify gaps. Strengthen resilience.

Book a Meeting



CREST Accredited

certin Empanelled

sales@briskinfosec.com