TALVIEW

AI Threat Index Report

The State of Exam Integrity in the Age of Generative AI

88%
Test-Takers Using AI
In assessed work globally, 2025

11×
AI Capability Growth
On expert-level exams, 2024–2025

94%
AI Work Undetected
Using standard detection tools

Three Structural Failures
Exam formats AI can complete without effort · Detection-only proctoring · Governance frameworks that cannot defend decisions to regulators or test-takers.

Six Attack Vectors Mapped
Agentic AI bypass · Remote desktop control · Earpiece relay · AI-accelerated braindumps · Deepfake identity substitution · Academic AI writing.

Trust as Infrastructure
Comprehensive 7-layer trust architecture designed for the agentic AI era, defensible to regulators and courts.

Data-Driven Intelligence
Verified incident data · Institutional disclosures · Live survey data from 9 assessment security leaders at the Talview Exam Security Summit 2026.

Executive Summary

Online examinations, professional certifications, and licensing assessments are in the middle of an integrity crisis the assessment industry was not designed to handle. The rapid democratization of generative AI tools — freely available, genuinely capable, and easy to use — has created a threat environment that outpaces the defensive infrastructure most organizations currently operate.

This report examines the AI-driven shift in exam fraud across higher education, professional certification, healthcare and legal licensing, and government-administered examinations. It draws on verified incident data, institutional disclosures, global survey research, and direct perspectives from nine assessment security practitioners who contributed at the Talview Exam Security Summit 2026. Coverage runs from October 2023 through March 2026.

88%
Test-Takers Using AI
In assessed work globally, 2025

11×
AI Capability Growth
On expert-level exams, 2024–2025

94%
AI Work Undetected
Using standard detection tools

The findings point to three structural failures that must be addressed: exam formats that AI can complete without meaningful effort; proctoring architectures designed for a pre-AI threat model; and governance frameworks that cannot defend AI-driven security decisions to regulators or test-takers.

The report closes with prioritized recommendations for exam administrators, certifying bodies, and their technology partners. The conclusion is not that digital assessment is broken — it is that the window to act is short and getting shorter.

CHAPTER 1 — THE THREAT HAS CHANGED

The Evolving Threat Landscape

Cheating in examinations is not new. What has changed is its architecture, its accessibility, and its scale. Historically, misconduct required deliberate coordination — purchasing answer keys, hiring a proxy, smuggling printed material. These approaches were costly, detectable, and limited in their reach. The emergence of large language models has eliminated all three constraints.

Any candidate with a smartphone and a free account can now receive a complete, fluent, contextually appropriate answer to almost any knowledge question in seconds. Detection is harder because AI-generated responses are indistinguishable from strong human work without specialist analysis — and even those tools are imperfect. Research confirms that 94% of AI-generated academic writing goes undetected by instructors relying on standard tools alone.

⚡ AI IS NOT A FEATURE UPGRADE. IT IS AN ADVERSARY UPGRADE.

For years, AI was positioned as a tool that could help exam security — better analytics, smarter anomaly detection, faster review. That framing is now obsolete. AI has crossed a threshold where it is no longer an aid to either side — it is the adversary. The security architecture must be built accordingly.

"

When assistance becomes instant, invisible, and indistinguishable, trust collapses quietly.

— Sanjoe Tom Jose, Chief Executive Officer (CEO), Talview — Exam Security Summit 2026 Keynote

INSTANT
Answers in seconds. No effort. No delay. No evidence of the time it should take.

INVISIBLE
Background applications, OS overlays, earpieces. No behavioral signature.

INDISTINGUISHABLE
Output quality matches or exceeds strong human work. Standard tools cannot tell the difference.

What makes AI-enabled cheating categorically different from every previous form of exam misconduct is captured in three words. AI assistance is now instant — answers arrive in seconds with no candidate effort. It is invisible — operating through background applications, OS-level overlays, and earpiece relay without any behavioral trace. And it is indistinguishable — the output quality of LLM-generated responses is, for most exam formats, identical to strong human work.

94%
AI Work Goes Undetected
Standard detection tools only

88%
Students Using AI
In assessed work globally, 2025

Three Generations of AI-Enabled Cheating

The threat has evolved through distinct phases since 2022. Each generation represents a structural escalation — not just in sophistication, but in the accessibility and invisibility of the attack.

01 LATE 2022
Direct Submission

In the first phase — coinciding with the launch of ChatGPT — test-takers primarily copied exam questions into publicly available AI tools and submitted responses verbatim. Detection relied on stylistic inconsistency and statistical score anomalies.

02 MID-2023 ONWARDS
Multi-Device Relay

More sophisticated techniques emerged: AI tools accessed via secondary devices outside camera sightlines, earpiece-and-relay systems where an external party processes questions and feeds answers in real time, and AI used to rapidly generate and distribute braindump content at speeds previously impossible.

CURRENT THREAT

03 2025 — PRESENT
Agentic AI — Current Phase

The most serious. AI agents can now operate autonomously within exam environments, completing questions, managing timing, and mimicking human interaction patterns without meaningful candidate involvement. The behavioral signals that proctoring systems were built to detect — gaze aversion, lip movement, typing cadence — are increasingly irrelevant when the candidate is not the one completing the exam.

⚡ AI IS NOT A FEATURE UPGRADE. IT IS AN ADVERSARY UPGRADE.

The framing matters. For years, AI was positioned as a tool that could help exam security — better analytics, smarter anomaly detection, faster review. That framing is now obsolete. AI has crossed a threshold where it is no longer an aid to either side — it is the adversary. The security architecture must be built accordingly.

A STUDENT CAN FALL ASLEEP...

\"A student can fall asleep. And amazingly, the computer is still populating answers. Because on the back end, they've set up code — some quite sophisticated, some very simple — that allows for bypass of monitoring systems.\"

— Jarret M. Dyer, Vice President of Test Security, Surpass Assessment

Evolving AI with Emerging Threats in AI Cheating Technologies

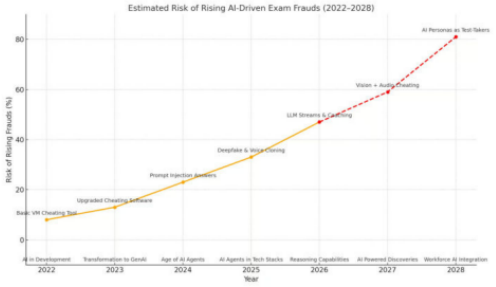
Year-by-year escalation of AI-driven exam fraud: from basic VM tools to AI personas replacing human test-takers entirely.

YEAR	EVOLUTION OF THE AGE OF AI	EMERGING THREATS IN AI CHEATING TECHNOLOGIES	RISK OF RISING FRAUDS (EST.)
2022	AI in Development	Basic VM Cheating Tool	8%
2023	Transformation to GenAI	Upgraded Advanced Cheating Software	13%
2024	Age of AI Agents	AI-Generated Answers with Prompt Injection	23%
2025	AI Agents Integrating in Tech Stacks	Deepfake Scams and Voice Cloning Tools	33%
2026	Advancements in Reasoning Capabilities	LLM-powered Answer Streams & Real-time Coaching	47%
2027	AI-Powered Real-Time Discoveries	Multimodal Cheating (Vision + Audio Assist Technology)	59%
2028	Integrate in Workforce Transformation	AI Personas Passing as Human Test-Takers	81%

Figure — Evolving AI Cheating Technologies & Estimated Risk of Rising Frauds (2022–2028). Talview Exam Security Intelligence.

Estimated Risk of Rising AI-Driven Exam Frauds (2022–2028)

Risk trajectory by cheating technology era — solid = confirmed trend · dashed = projected escalation



⚠ 8% → 81% IN SIX YEARS

The risk of AI-driven exam fraud is projected to reach **81%** by 2028 as AI personas become capable of passing as human test-takers with no detectable behavioral fingerprint. Organizations that do not adopt trust infrastructure before this threshold will face structural credibility failure.

The Pace of Capability Improvement

What makes this moment categorically different from previous cycles of exam fraud is the rate at which the underlying technology improves.

AI Capability on Expert-Level Academic Tests

The trajectory that should define every security decision · Humanity's Last Exam benchmark · % correct

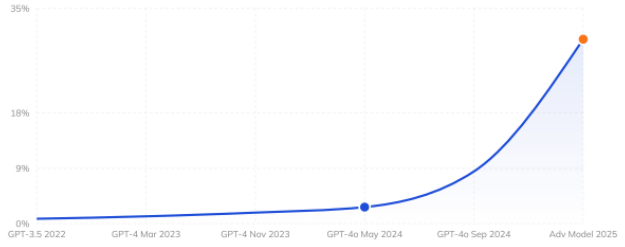


Figure 1 — AI model performance on expert-level academic benchmarks. GPT-4o score on Humanity's Last Exam, 2024 vs 2025.

SINGLE-YEAR IMPROVEMENT

11×

GPT-4o → Advanced Model
2024 to 2025

GPT-4o achieved a **2.7% correct rate** in 2024. By 2025, a more advanced model reached **30%** on the same benchmark — an 11× improvement in a single year. Applied to the question types that appear in most professional and academic examinations, this trajectory means the window for detection-based defense is narrowing faster than most organizations are adjusting.

"The pace of improvement is quite striking. The speed at which the technology is improving creates particular challenges in the assessment space — it always feels like a bit of an arms race."

— Dan Hughes, Vice President of Product Development, Talogy

Six Attack Vectors — By Severity

The threat landscape spans six primary attack vectors, varying in how common they are, how serious their downstream impact is, how hard they are to detect, and how fast they are spreading.

AI Threat Severity Index by Attack Vector (2025)

Composite score: Prevalence × Impact × Detectability Difficulty × Proliferation Velocity

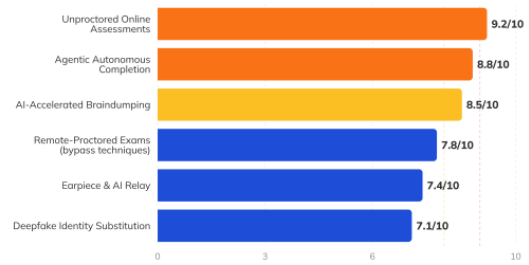


Figure 2 — AI Threat Severity Index by Attack Vector (2025). Composite score: Prevalence × Impact × Detectability Difficulty × Proliferation Velocity.

Severity Classification

- Critical (9.0–10)
- High (8.0–8.9)
- Elevated (7.0–7.9)

⚠️ The Critical Insight

The most important insight is not the top score — it is the score for **remote-proctored examinations: 7.8/10**. Remote proctoring is the industry's primary response to exam fraud. The data says sophisticated bypass techniques including remote desktop control, screen virtualization, and earpiece relay remain largely undetectable by current-generation proctoring solutions.

9.2	Unproctored Online Assessments	Critical
8.8	Agentic Autonomous Completion	Critical
8.5	AI-Accelerated Braindumping	High

AI Usage Is Already Widespread — Across Every Sector

AI Tool Usage Among Exam Test-Takers by Sector

2024 vs. 2025 · Sources: HEPI (2025), BestColleges, Turnitin, nerdynav · % of test-takers using AI in assessed work

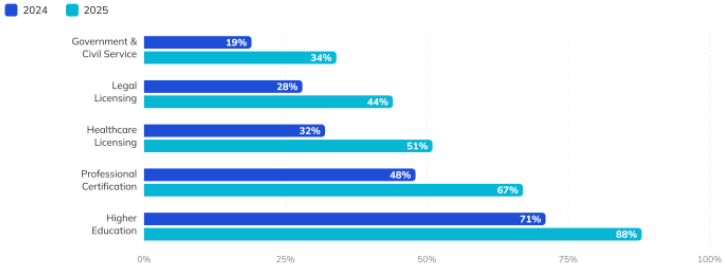


Figure 3 — AI Tool Usage Among Exam Test-Takers by Sector. Sources: HEPI (2025); BestColleges, Turnitin, nerdynav.

Higher education leads at **88%** by 2025. The steeper risk lies in professional certification, healthcare, and legal licensing — domains where credential fraud carries direct public safety implications and detection infrastructure is least developed.

Year-over-Year Change

Higher Education
71% → **88%** +17pp

Professional Certification
48% → **67%** +19pp

Healthcare Licensing
32% → **51%** +19pp

Legal Licensing
28% → **44%** +16pp

Government & Civil Service
19% → **34%** +15pp

⚠️ PUBLIC SAFETY RISK Healthcare and legal licensing represent the highest public safety risk. A clinically unqualified practitioner or legally undertrained professional who passed through AI assistance creates real-world harm that extends far beyond the examination event.

Detection Is Dead

The assessment industry's predominant response to AI-enabled cheating has been detection. The comforting logic: if we can detect an action, we can fix this.

🔵 THE COMFORTING MYTH

"If we can detect an action, we can fix this." — The assumption behind every detection-first architecture deployed in the last three years.

More sophisticated anomaly scoring, AI-generated content detectors, statistical outlier flagging. Build enough detection, the thinking goes, and the problem becomes manageable.

🔴 THE HARSH REALITY

- Humans cannot reliably detect AI-generated work at scale
- AI detectors produce systematic false positives against legitimate test-takers
- False negatives — undetected fraud — go unnoticed and unchallenged
- Flags are not enforcement: a flag without a defensible decision changes nothing

DETECTION PRODUCES NOISE. INFRASTRUCTURE PRODUCES OUTCOMES.

Every detection-only system operates on the same flawed assumption: that identifying a suspicious event creates a defensible outcome. It does not. A flag is not a finding. An alert is not evidence. A suspicion score is not a decision that can withstand a candidate appeal, a regulatory audit, or a legal challenge. The organizations building the most defensible security programs have stopped optimizing their detection layer — and started building infrastructure.

False Positive Problem

AI detectors trained on general corpora systematically flag ESL writers, neurodivergent test-takers, and high-achieving students as suspicious — eroding test-taker trust before enforcement even begins.

False Negative Problem

The 94% of AI-assisted work that goes undetected by standard tools creates a selection effect: only the most unsophisticated or unlucky test-takers get caught, while organized operations remain invisible.

Defensibility Gap

A flag cannot be defended to a court, a regulator, or a test-taker's appeal without a specific, documented, human-reviewed decision chain. Detection without infrastructure is theater, not security.

When It Actually Happened — Part 1

The following cases represent the most significant AI-related exam integrity failures documented during the coverage period.

ORGANIZATION	INCIDENT SUMMARY	SEVERITY	IMPACT
ACCA Professional Accounting	340% surge in AI-assisted cheating (Q1 2024–Q3 2025). Test-takers used LLMs while maintaining normal behavioral patterns — making eye-tracking and keystroke analysis obsolete. CEO cited sophistication "far outpacing available safeguards." Remote exams terminated March 2026.	CRITICAL	Full rollback of remote examination model affecting hundreds of thousands globally. Loss of disability accommodations, cost savings, and remote access. The clearest proof that detection-only proctoring fails at scale.
KPMG Netherlands CPA / Professional Licensing	PCAOB fined KPMG Netherlands USD 25M for systematic professional licensing exam cheating and deliberately misleading regulators. Disclosed April 2024; conduct spanned multiple cohorts and organized answer-sharing across exam sittings.	CRITICAL	Record regulatory penalty. Exposed the structural vulnerability of self-administered corporate licensing programs. Triggered cross-firm regulatory scrutiny of professional accounting credentials globally.
Ernst & Young USA CPA Ethics Exam	SEC imposed USD 100M fine for widespread cheating on CPA ethics components and mandatory professional licensing courses over multiple years. Settlement also reached with Connecticut state regulators.	CRITICAL	Largest penalty of its type. Established precedent for cascading regulatory consequences. Reframed credential fraud from an individual-conduct issue to a systemic professional-services risk.
USMLE Medical Licensing Nepal Recall Ring	Statistically impossible cluster of 99th-percentile scores linked to a small number of Nepali medical schools uncovered January 2024. Organized recall-bank operation sold exam content for USD 300–2,000 per package. AI now accelerates harvesting and distribution.	CRITICAL	Score invalidation for multiple test-takers. Direct public safety risk from clinically unqualified practitioners entering healthcare. Exposed the commercial recall-marketplace now dramatically scaled by AI.
California State Bar 2025 Bar Examination	29 of 200 MCQs were generated by AI without disclosure to test-takers, the legal community, or the state Supreme Court. Exam delivery platform simultaneously collapsed. Fraud lawsuit filed against the delivery vendor.	CRITICAL	Introduced a new risk category: AI misuse within exam administration itself. Combined content-validity and delivery failures caused mass test-taker disruption and eroded public confidence in bar admissions.

When It Actually Happened — Part 2

ORGANIZATION	INCIDENT SUMMARY	SEVERITY	IMPACT
Microsoft Certifications Azure / Security / M365	328 active braindump sites identified selling Microsoft exam content. New exam questions appearing online within 24 hours of first release. AI dramatically accelerates content harvesting, organization, and commercial distribution.	HIGH	IT certification market flooded with credential holders lacking genuine competence. Employer reliance on certifications as hiring signals materially eroded. Forced accelerated item rotation cycles.
UK Universities 131 Institutions	FOI disclosures: approximately 7,000 confirmed AI cheating cases in 2023–24 — 5.1 per 1,000 students, triple the prior year rate. LSE reported a 468% spike in detected AI misconduct. Researchers found 94% of AI-submitted work went undetected via standard tools.	HIGH	Estimated investigation cost: £12.4M annually in the UK, USD 196M in the US. Actual prevalence significantly higher — institutions with stronger detection consistently report proportionally more cases.
NAATI Language Interpreter Exams	Test-takers discovered wearing wigs concealing AirPods with live AI translation during interpreting assessments — administered in 98 languages, all online. A new class of hybrid threat: physical concealment combined with AI real-time assistance.	HIGH	Direct public safety implications in medical, legal, and emergency contexts. Detection requires new approaches combining behavioral, environmental, and device monitoring simultaneously.

THE PATTERN WORTH NAMING

The EY and KPMG cases establish that AI-assisted credential fraud is not a candidate-conduct problem — it is a **systemic professional-services risk** with regulatory, legal, and market-confidence consequences that extend years beyond individual exams. India's 2025 legislation is the first national framework globally to respond at that scale. More will follow.

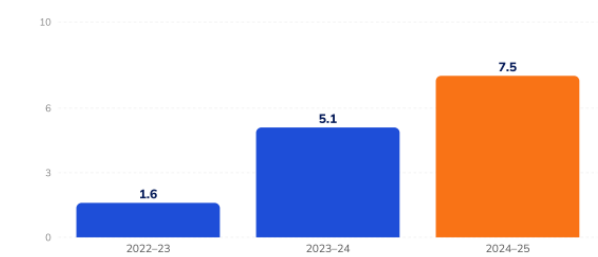
Across these cases, three structural patterns recur: **organized commercial actors** rather than isolated individuals driving the most serious incidents; detection-only proctoring consistently failing against AI tools that generate no traditional behavioral signals; and downstream consequences — financial, regulatory, public safety — **extending far beyond the examination event itself**.

UK University AI Cheating: Three-Year Escalation

FOI survey of 131 universities (Source: The Guardian, 2025)

Proven AI Cheating Cases per 1,000 Students

2022–2025 · FOI survey of 121 universities (Source: The Guardian, 2025)



Confirmed · 2024–25

219% Increase 2022–23 to 2024–25
From 1.6 to 7.5 per 1,000 students

Why These Numbers Are An Undercount

The UK data is instructive precisely because it reflects only **detected cases**. With 94% of AI-submitted work going undetected via standard tools, the 7,000 confirmed UK cases in 2023–24 are almost certainly a fraction of actual total — a conclusion consistent with the observation that **institutions investing more in detection report proportionally more cases**.

7,000 Confirmed UK cases
2023–24 academic year

£12.4M Annual investigation cost
UK universities annually

$196M US investigation cost est.
Annual estimate

468% AI misconduct spike
LSU reported increase

AI Threat Index & Trend Data

Every measure points the same direction. The following data synthesizes incident trajectory from institutional disclosures and survey research.

CHATGPT 4O, 2024–2025	SEVERITY INDEX SCORE	UNIVERSITY OF READING STUDY	KPMG + EY COMBINED
11×	**9.2/10**	**94%**	**$125M+**
AI Capability Growth	Unproctored Exam Risk	AI Work Undetected	Regulatory Fines

AI-Assisted Cheating: Indexed Incident Growth by Exam Category

2022–2025 · Base = 100 at Q1 2022 · Key inflections: ChatGPT launch Nov 2022; GPT-4 Multimodal Mar 2024

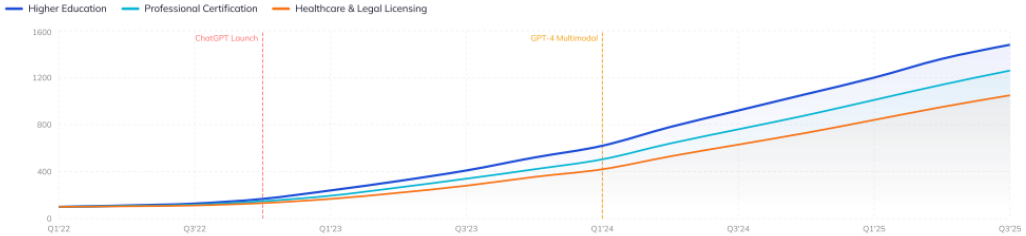


Figure 6 — AI-Assisted Cheating: Indexed Incident Growth, 2022–2025. Base = 100 at Q1 2022.

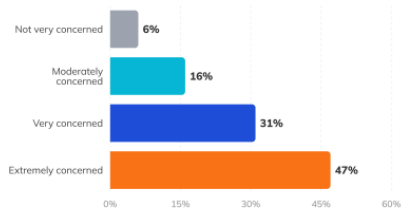
Live Survey Data — Talview Exam Security Summit 2026

When nine assessment security leaders gathered at the Summit, they presented what is happening in their programs right now.

SURVEY 1

How concerned is your organization about AI-enabled cheating in the next 12 months?

Live survey — Exam Security Summit 2026 · (n = summit practitioners)

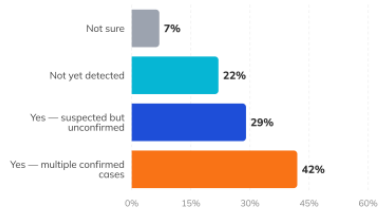


78% **Extremely or Very Concerned**
About AI-enabled cheating in the next 12 months

SURVEY 2

Has your organization already detected AI-assisted cheating incidents?

Live survey — Exam Security Summit 2026



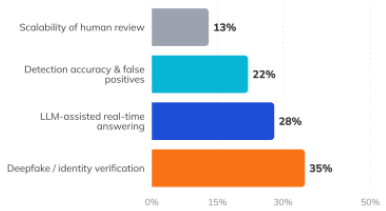
71% **Already Detected Incidents**
Confirmed or suspected AI cheating in active programs

Live Survey Data — Top Challenges & Key Findings

SURVEY 3

What is your biggest exam security challenge today?

Live survey — Exam Security Summit 2026



35% **#1 Challenge: Deepfake / Identity Verification**
The single biggest exam security challenge identified by practitioners

📊 WHAT THE SURVEYS TELL US

78% of summit practitioners are 'extremely' or 'very' concerned about AI-enabled cheating in the next 12 months. Over 70% have already detected suspected incidents in their programs. The top two challenges — deepfake identity verification and LLM-assisted real-time answering — both point to the same gap: the proctoring architecture was built for a world where a candidate's physical presence and their exam performance were reliably connected. **That connection can now be severed.**

🎭 **Identity Cannot Be Assumed**
Deepfakes and identity fraud are now the #1 challenge. The question has shifted from "is this the right person?" to "is this a real person at all?"

🤖 **Real-Time AI Is the Core Threat**
LLM-assisted answering during the exam session represents a threat that operates without any behavioral signature visible to current proctoring systems.

⚖️ **Human Review Does Not Scale**
Scalability of human review ranks as a challenge because detection generates more flags than organizations can review — and a flag without human review is meaningless.

What the Industry Is Experiencing: The Discreteness Problem

The hardest-to-detect cheating today leaves no obvious behavioral signature.

The Discreteness Gap: Why Traditional Proctoring Can No Longer See the Threat

WHAT THE PROCTOR SEES	Detection Gap	WHAT'S ACTUALLY HAPPENING
👀 — Candidate sitting calmly		🤖 — AI agent completing exam autonomously
👁️ — Eyes on screen		💻 — OS-level overlay reading questions
⌨️ — Normal typing cadence		🎧 — LLM answers fed via Bluetooth earpiece
📱 — No secondary device visible		🖥️ — Remote desktop operator in control
📊 — Behavioral score: Normal		⚠️ — Risk score: Invisible to current systems

Figure 7 — The Discreteness Gap: what proctors see versus what is actually happening in current AI-assisted cheating scenarios.

> *A student can fall asleep. And amazingly, the computer is still populating answers. Because on the back end, they've set up code — some quite sophisticated, some very simple — that allows for bypass of monitoring systems.*
>
> — Jarret M. Dyer, Vice President of Test Security, Surpass Assessment

> *"The benefit we have at the moment — our advantage — is rapidly eroding as these technologies improve. The speed and the discreteness of it is what concerns me most."*
>
> — Michael Nemarich, Chief Operating Officer, NAATI

📈 **Performance Anomalies as Signal**

Performance anomalies are becoming the most reliable signal available: a candidate scoring the 19th percentile and then the 99th percentile between consecutive sittings with no detectable behavioral change. These reflect outcomes rather than behaviors — and they are harder to argue away.

Cheating-as-a-Service: When Fraud Becomes a Business

One of the most significant structural shifts in the fraud landscape is the commercialization of AI-assisted cheating into subscription and service-based products.

These are not informal peer networks — they are organized businesses offering guaranteed exam performance, built on AI overlay technology that intercepts the display output of secure browsers, processes visible question content, and relays generated answers in real time.

How Cheating-as-a-Service Operates: The Commercial AI Fraud Pipeline

1. **Candidate Registers** — Pays subscription or per-exam fee to CaaS provider
2. → **CaaS Provider Activates System** — No personally identifiable info required; everyone deployed
3. → **Candidate Strategy: Screen intercepted** — Questions extracted via OS-level overlay or screen capture
4. → **LLM Processes Questions** — Real-time AI processes exam content, generates answers
5. → **Answers Delivered to Candidate** — Bluetooth audio; screen overlay; AI-dictated typed response

Result: Candidate passes. Provider profits. Exam integrity fails.

Figure 8 — How Cheating-as-a-Service operates: the five-step commercial AI fraud pipeline.

> *There are certain applications — I don't want to give names — but they sell cheating as a service. Organizations making money out of a cheating system you can plug into secure exam delivery engines. An AI-assisted overlay screens between the proctoring system and the student, reading what is shown and prompting what kind of replies to give.*
>
> — Miguel Prieto, Vice President of Corporate Strategy, Open Assessment Technologies

⚠️ **Secondary Consequence: Extortion**

Test-takers who pay for CaaS services sometimes face extortion from the providers they hired — paying once to pass, and again to avoid exposure. This transforms exam fraud from a compliance problem into broader criminality that extends well beyond the exam function.

A Three-Tier Autonomy Framework for Trustworthy AI in Exam Security

The most practically useful governance framework divides AI decisions in examinations into three tiers based on consequence. The distinction between tiers is not technical — it is about accountability.

TIER 1

AI DECIDES — Out of the Loop

Human awareness: Low-consequence, AI-autonomous

Low-stakes, reversible checks. AI acts autonomously with full audit trail. No human review required unless overridden.

TYPICAL DECISIONS

Biometric check | Camera test | System readiness | ID pre-verification

AI Automation Level **95%**

TIER 2

AI ACTS — Human Monitors (On the Loop)

Real-time anomaly flagging; human can overrule at any moment

Consequential decisions. AI packages evidence; qualified human reviewer monitors and can override. Alert without review is unacceptable.

TYPICAL DECISIONS

Secondary audio | Screen activity | Score pattern alerts

AI Automation Level **60%**

TIER 3

HUMAN DECIDES — AI Provides Evidence (In the Loop)

AI mandates this for high-stakes educational decisions

Human makes the final call. AI provides all evidence and reasoning but cannot act autonomously. Required for high-stakes decisions.

TYPICAL DECISIONS

Disqualification | Score invalidation | Regulatory disclosure | Credential revocation

AI Automation Level **10%**

CRITICAL GOVERNANCE RISK

⚠️ The commercial pressure to collapse tier-three decisions into tier-one automation — eliminating human review to reduce cost — represents the industry's most significant near-term governance risk. Even at Microsoft, leadership was reluctant to allow AI-driven proctoring decisions on certification test-takers until transparency requirements around training data and decision logic were met.

CHAPTER 5 — THE DEFENSIBILITY IMPERATIVE

The Defensibility Imperative

The ACCA decision to terminate remote examinations in early 2026 represents a watershed moment — but its significance lies not in the reversion to in-person testing. It lies in what that decision reveals: organizations have deployed AI monitoring systems **without building in the governance, transparency, and human-oversight structures** needed to defend their decisions to test-takers, regulators, and courts.

THE HIGH-STAKES REALITY

- When a license is granted — a medical practitioner cleared to treat patients, an engineer certified to sign off on structures.
- When a certification is earned — a financial adviser authorized to manage retirement savings, a security professional trusted with critical infrastructure.
- When a job is offered — based on a credential that an AI completed without the candidate's meaningful involvement.

In each of these moments, you do not need more alerts. You need defensible decisions.

Is our AI a guard dog that prevents intrusions, or is it a scarecrow that looks intimidating but can't actually stop anything? The illusion of control happens when you have a dashboard full of red flags but no actual reduction in malpractice. You're not proctoring the exam anymore — you're proctoring the AI's alerts.

— Sundar Nagarathnam, Strategic Advisor

What Trust as Infrastructure Means

Infrastructure is different from a feature in four critical ways. These are the properties that payment systems, identity platforms, and cybersecurity architectures share — and that exam security has not yet adopted at scale.

ASSUMES FAILURE

Designed with resilience in mind. Does not depend on any single layer being intact. Built for the expectation that individual controls will fail.

LAYERS DEFENSE

Multiple independent security measures. A failure in one does not expose the whole. Seven reinforcing layers, each independently defensible.

PRODUCES PROOF

Generates verifiable evidence and auditable trails — not alerts that fade. Every consequential decision has a documented evidence chain.

SURVIVES SCRUTINY

Robust against candidate challenges, regulatory audits, and legal examination. Designed to answer the question: can we defend this decision?

"This is how payments work. This is how identity works. This is how security works. Assessment integrity is next."

— Sanjoe Tom Jose, Chief Executive Officer (CEO), Talview — Exam Security Summit 2026 Keynote

CHAPTER 5 — THE DEFENSIBILITY IMPERATIVE

Conditions for a Trustworthy AI Security Decision

"AI is only trustworthy when the evidence is clear, the reasoning behind the flag is explainable, the action aligns with the policy, a human is in the loop for high-stakes decisions, and the response is proportionate to the risk. We often overreact — and sometimes mistake normal test-anxiety behavior for fraud."

— Liberty Munson, Director of Psychometrics, Microsoft

01 Explainability

The system must produce a specific, human-readable account of what triggered a flag — not "an anomaly was detected" but "the candidate directed their gaze toward a secondary device for 47 seconds across three separate intervals."

02 Traceability

Evidence must be reviewable by a qualified human reviewer independently of the AI that generated it. The reviewer must be able to reconstruct the decision without relying on the AI's own explanation.

03 Policy Alignment

The AI's decision logic must map to written, candidate-disclosed policies. If the AI flags a behavior that no published policy prohibits, the flag cannot form the basis of a disciplinary action.

04 Proportionality

The severity of automated response must correspond to the confidence level of the flag. High-confidence flags with clear evidence may support automated alerts. Low-confidence flags must not trigger automated disqualification.

⚠ THE RISK OF GETTING THIS WRONG AT SCALE

The biggest risk isn't just technical failure. It's what I call scalable injustice. Think about what happens when autonomous agents can flag, block, or disqualify test-takers at scale — but without meaningful human oversight. You can suddenly wake up and find that a system has incorrectly flagged hundreds of honest test-takers overnight. The damage is not just to the individuals — it's the credibility of the examination system itself.

— Sundar Nagarathnam, Strategic Advisor

How AI Systems Fail Differently

Traditional systems that fail do so slowly — individual cases, individual reviews, individual appeals. **AI systems fail at the speed and scale of automation.** A miscalibrated flag threshold does not produce one wrongful disqualification. It can produce hundreds overnight.

CHAPTER 6 — PREVENTION: WHAT WORKS, WHAT DOESN'T

What Is Working

The current evidence base points clearly to what categories of intervention are effective. No single control is sufficient; the most resilient programs operate across multiple defensive layers simultaneously.

01 AI-Resilient Exam Design

Applied and scenario-based formats requiring contextual judgment produce responses that AI generates poorly — and that trained examiners can evaluate. Adaptive question pools with high rotational depth defeat braindump utility. Oral components, performance simulations, and time-pressured tasks requiring environmental context remain the most AI-resistant formats available.

"People who are already designing their assessments to be enriched with AI will be in a better position than those trying to combat it. The key is to adapt to new technologies — evolving assessment design is going to be the defining strategic choice."

— Miguel Prieto, Vice President of Corporate Strategy, Open Assessment Technologies

Oral components | Performance simulation | Adaptive pools | Time-pressured tasks

2 Cross-Session Pattern Analytics

Moving from single-session anomaly detection to cross-session pattern analysis is the most operationally significant upgrade available without hardware changes. Test-takers moving from the 19th to the 99th percentile between attempts, IP clustering, and geographic concentration analysis remain the most actionable signals. Detecting patterns, not events, changes what is findable.

Percentile jump detection | IP clustering | Geographic analysis | Cross-session scoring

3 Zero-Trust Baseline Design

Start from the assumption that AI can complete everything in your exam, then design defensively from that baseline. When controls fail that test, they are redesigned before deployment — not after an incident. This is the principle that NAATI operates on across 98 languages of live examination.

Assume AI can answer everything | Design defensively | Pre-deployment testing

4 Continuous Biometric Authentication

Multi-factor identity assurance maintained throughout the session — not only at login — using behavioral biometrics calibrated to a per-candidate baseline. Liveness detection that goes beyond static photo matching is essential to address deepfake substitution.

Continuous auth | Per-candidate baseline | Liveness detection | Deepfake defense

CHAPTER 6 — PREVENTION: WHAT WORKS, WHAT DOESN'T

What Is No Longer Sufficient

These controls have passed their useful operational life against the current threat. They may remain as one input among many — but cannot anchor a security program.

Detection-Only Proctoring Architectures

Obsolete as primary defense

Eye-tracking, keystroke analysis, and ambient audio monitoring were not designed for test-takers using AI on separate devices with no behavioral fingerprint. These signals remain useful as one input among many but cannot be the primary defense. The ACCA case is the definitive 2025 demonstration of their limitations at scale.

Eye-tracking | Keystroke analysis | Audio monitoring

Static Question Banks of Finite Depth

Structurally compromised

Any bank that can be fully enumerated can be systematically harvested. Continuous rotation and adaptive generation are necessary in high-volume programs. Organizations operating on annual item refresh cycles face material exposure.

Fixed MCQ banks | Annual refresh cycles | Low-depth pools

Honor Agreements as Primary Deterrents

Insufficient alone

Signed attestations are a necessary component of the integrity framework but carry no deterrent effect against test-takers who have already decided to use AI assistance. Credible enforcement consequences must accompany policy disclosures to produce behavioral change.

AI policy agreements | Signed attestations

Effectiveness Comparison

Pre-agentic AI era vs. now (0–100 scale)

	Was	Now
AI-Resilient Design	45	88
Cross-Session Analytics	30	82
Zero-Trust Baseline	25	78
Continuous Biometrics	55	84
Detection-Only Proctoring	72	22
Static Question Banks	65	18
Honor Agreements Only	50	8

Key insight: Controls designed for pre-AI cheating have dropped sharply in effectiveness, while AI-resilient approaches have surged. Organizations still anchored to old controls face compounding risk.

CHAPTER 7 — ACTING BEFORE THE WINDOW CLOSES

Immediate Actions — 0 to 6 Months

The evidence base in this report does not support pessimism about the future of digital assessment. It supports urgency.

The organizations maintaining stable credentialing integrity through the AI transition share common characteristics: they treat security as a continuous program rather than a compliance checklist; they invest in practitioner training alongside technology; and they have built governance structures capable of defending their decisions to external scrutiny.

⏱ IMMEDIATE PRIORITY · 0–6 MONTHS

ACTION 1

Audit your threat exposure

Map your current assessment portfolio against the AI Threat Severity Index. Identify which formats — particularly unproctored or lightly monitored digital assessments — carry critical-tier risk under current controls.

ACTION 2

Update candidate AI policy with specificity

Revise examination terms to explicitly prohibit defined categories of AI tool use, with consequences that are proportionate, published in advance, and consistently enforced. Honor agreements without enforcement produce no deterrent effect.

ACTION 3

Add post-exam AI content analysis

Implement stylometric and AI-content detection as standard quality assurance for all remotely delivered high-stakes examinations. This provides a secondary layer that many current architectures entirely lack.

ACTION 4

Shift from detection to deterrence

Transparent communication of monitoring capabilities — what the system observes, what constitutes a violation, what consequences follow — reduces the population willing to attempt AI assistance before the session begins.

CHAPTER 7 — ACTING BEFORE THE WINDOW CLOSES

Strategic Priorities — 6 to 18 Months

🎯 STRATEGIC PRIORITIES · 6–18 MONTHS

PRIORITY 1

Redesign highest-risk exam formats

Begin systematic replacement of multiple-choice and short-answer components in high-value credentials with applied, performance-based, or adaptive formats. This is the only permanent solution — technology controls are indefinitely outpaced by AI capability improvements.

PRIORITY 2

Build AI governance into procurement

Evaluate proctoring and security solutions against defensibility criteria: decision explainability, human-override capability, auditable evidence trails, proportionality of automated actions — not detection accuracy alone.

PRIORITY 3

Establish cross-organization intelligence sharing

Cheating syndicates, braindump operations, and CaaS providers operate across organizational and jurisdictional boundaries. Coordinated legal action and threat intelligence sharing require equivalent cross-boundary infrastructure.

PRIORITY 4

Engage regulators proactively

Contribute to policy frameworks under the EU AI Act's high-risk education classification, India's 2025 legislative model, and equivalent national frameworks. Organizations that shape these frameworks will have more defensible compliance postures.

TRUST AS INFRASTRUCTURE — BUILDING THE DEFENSE THE THREAT DEMANDS

Talview's 7-Layer Trust Infrastructure

The first architecture of its kind in the assessment industry. Patented (US Patent No. 12,361,115 B1, granted July 2025).

#	LAYER	WHAT IT DOES	THREAT ADDRESSED
1	**Identity Verification**	AI-aided document checking, real-time liveness detection, deepfake-specific detection models, environment validation	Identity fraud & impersonation
2	**Secure Browser (OS-level)**	OS-level app detection blocking AI overlays, remote desktop/VM detection, screen protection and minimizing enforcement	Screen capture & overlay bypass
3	**Behavioral Biometrics**	Per-candidate baseline calibration, continuous authentication throughout session, typing cadence and interaction patterns	Agentic AI impersonation
4	**Session Monitoring**	Real-time anomaly scoring, gaze analysis, environmental monitoring via secondary camera for 360° scanning	Earpiece & relay systems
5	**Content Integrity Analysis**	Post-exam stylometric analysis, AI-generated content detection, response pattern anomaly review	AI-generated responses
6	**Cross-Session Intelligence**	Pattern analysis across exam sittings, IP/geographic clustering, score percentile jump detection, collusion mapping	Organized braindump networks
7	**Human Oversight & Governance**	Tier-2 and Tier-3 escalation framework, audit-ready evidence packaging, proportionate response protocols	Scalable injustice & defensibility

Layer Coverage by Threat Vector

- Layer 1: Identity Verification — Active
- Layer 2: Secure Browser (OS-level) — Active
- Layer 3: Behavioral Biometrics — Active
- Layer 4: Session Monitoring — Active
- Layer 5: Content Integrity Analysis — Active
- Layer 6: Cross-Session Intelligence — Active
- Layer 7: Human Oversight & Governance — Active

THE CORE PRINCIPLE

Not one model. Not one flag. Seven reinforcing layers — each independently defensible, together forming a system no single attack vector can defeat.

"AI created the problem. Agentic AI enforces the solution."

— Sanjoe Tom Jose, Chief Executive Officer (CEO), Talview
— Exam Security Summit 2026 Keynote

TRUST AS INFRASTRUCTURE — TALVIEW PROCTORING V8

Talview Proctoring V8: The Agentic AI Trust Infrastructure

Where traditional proctoring generates events, V8 produces structured incidents. Where traditional systems issue alerts, V8 builds verifiable evidence.

TRADITIONAL PROCTORING	V8 ENFORCEMENT
✗ Generates events	✓ Produces structured incidents
✗ Issues alerts	✓ Builds verifiable evidence
✗ Creates suspicion	✓ Delivers defensible outcomes
✗ Flags for human review	✓ Packages cases for audit-ready action

FALSE POSITIVES ERODE CREDIBILITY. FALSE NEGATIVES DESTROY TRUST.

V8 is built to eliminate both failure modes. The system is designed not just to catch fraud — but to produce evidence that can withstand a challenge, and to clear legitimate test-takers without generating the false positive burden that has undermined confidence in AI proctoring systems industry-wide.

75% — Reduction in review times — Copilot mode

40% — Reduction in false alerts — Autopilot mode

99% — Events recognized vs humans — Fewer false positives

35% — Higher candidate satisfaction — Contextual assistance

Alvy — The World's First Patented Agentic AI Proctoring Agent

US Patent No. 12,361,115 B1, granted July 2025

INTERACT
Engages with test-takers, environments, and systems in real time — not after the fact.

REASON
Synthesizes signals from seven data layers into a coherent judgment — not a rulebook lookup.

ESCALATE
Routes cases requiring human judgment to the appropriate tier — proportionate, documented, defensible.

LEARN
Adapts continuously as new attack patterns emerge — making the system harder to defeat with each release.

CHAPTER 4 — THE GROUND-LEVEL REALITY

Who Loses Trust First — and What Is at Stake

When examination integrity fails, the question is who bears the consequences first.

Regulatory Bodies — HIGH

Most visibly affected when incidents become politically exposed

Reactive responses — banning specific tools, reverting to paper formats — address the symptom rather than the cause

Employers — MEDIUM

Erosion of confidence in credentials is slower but more durable

When a significant proportion of a certification pool is believed to have passed with AI assistance, employers reduce reliance on it as a hiring signal

Certification Bodies — HIGH

Credential market value declines. The consequences persist long after specific incidents are resolved

Must rebuild trust from scratch — which takes years and requires demonstrable systemic change

Honest Test-Takers — CRITICAL

"Guilty until proven human" — increasingly possible as AI tools proliferate

Legitimate test-takers bear the burden of proving authenticity rather than placing burden of proof on the accusing system

The Presumption of Guilt Risk

The deeper risk is more subtle. As organizations strengthen AI-based security infrastructure, there is a real possibility of drifting toward a **presumption of guilt** — an inversion of standard evidentiary principles that forces honest test-takers to prove their authenticity rather than placing the burden of proof on the accusing system.

The Market Confidence Cascade

1. Significant fraud event occurs (confirmed or suspected)
2. Employers reduce reliance on credential as hiring signal
3. Credential market value declines
4. Legitimate holders face diminished returns
5. Trust recovery takes years — requires systemic change

"Guilty until proven human." That is where we could end up. There are so many ways people can use agentic AI tools — one agent doing screen recalls, one solving problems, another rephrasing the answer. The defense shifts from "do we know they cheated?" to "how do we know a human was even there?"

— Paul Muir, Chief Customer Officer, risr/

Among employers, the erosion of confidence in credentials is slower but more durable. When a significant proportion of a certification pool is believed to have passed with AI assistance — proven or not — the market value of that credential declines.

When AI Becomes The Attacker, Trust Must Become Infrastructure.

88%	94%	11×	$125M+	7,000+	257%
Test-Takers using AI	AI work undetected	Capability growth	Regulatory fines	UK cases confirmed	Deepfake surge

> *AI created the problem. Agentic AI enforces the solution.*
>
> — Sanjoe Tom Jose, Chief Executive Officer (CEO), Talview — Exam Security Summit 2026 Keynote

> *Don't give up on it — but constantly evolve. Invest and constantly evolve. The threat is accelerating. The response must accelerate faster.*
>
> — Michael Nemarich, Chief Operating Officer, NAATI

> *When AI attacks trust, we shouldn't abscond. We should respond. And to respond, we need a trust infrastructure.*
>
> — Sanjoe Tom Jose, Chief Executive Officer (CEO), Talview

TRUST AS INFRASTRUCTURE — CORE PRODUCT SUITE

Talview Platform: Core Product Suite

Deepfake incidents surged 257% in 2024. The identity verification problem shifted from 'is this person who they say they are?' to 'is this a real person at all?'

Alvy — Agentic AI Proctoring Agent

- Copilot mode: Doubles proctor coverage, cuts review times by up to 75%
- Autopilot mode: Fully automated monitoring with 40% reduction in false alerts
- Detects and counters Cluely, FinalRound AI, virtual machines, LockedIn, screen overlay tools
- Recognizes 99% of events flagged by human proctors — with fewer false positives
- 35% higher candidate satisfaction through real-time contextual assistance
- AI-driven risk scoring using cross-session analytics to detect collusion and identity fraud

Secure Browser — OS-Level Defense

- OS-level app detection blocks Cluely-style AI overlays and screen interception tools
- Remote desktop and virtual machine detection — the bypass technique ACCA missed at scale
- Screen protection and minimizing enforcement — prevents content capture
- Available as browser extension for Moodle 4.5 and leading LMS platforms
- Two-factor authentication and Proview Index for pre-exam trust establishment

ID Verification — Defeating Deepfakes at the Gate

- Real-time face recognition and photo ID matching to prevent impersonation
- Deepfake detection models trained on synthetic media — not just static photo matching
- Liveness detection maintained throughout the session, not only at login
- Environment monitoring via secondary camera for 360° room and device scanning
- Continuous biometric authentication calibrated to per-candidate enrolled baseline

INTEGRATION ECOSYSTEM

LMS PLATFORMS
Moodle · Canvas · Skilljar · D2L · Intellum and more

ASSESSMENT ENGINES
risr/ · Surpass Assessments · RM Assessments · Inspera · iMocha and more

TALVIEW

Thank You

Exam integrity is not a technology problem — it's a trust infrastructure challenge. We hope this analysis provides a foundation for building that infrastructure.