CASE STUDY: HOW COSMOS AI AND HUMAN EXPERTISE WORK TOGETHER TO STRENGTHEN APPLICATION SECURITY

AT A GLANCE

0

False Positives Delivered
All findings confirmed by a Bishop Fox expert before delivery

3
Business Logic Flaws Found by AI

20
Confirmed Vulnerabilities

3H 17M
AI Scan Duration

THE CHALLENGE:

As part of a proof-of-value engagement, a financial services organization provided a test application to evaluate the capabilities of Bishop Fox's Cosmos AI security testing platform against a realistic target. The application was a Flask/Python platform simulating customer-facing functionality: user accounts, financial transactions, trading activity, and administrative functions. The target represented a realistic attack surface: multi-role authentication (standard user and administrator), 20+ API endpoints spanning account management, financial transfers, trading, and reporting, and server-side transaction logic typical of production financial applications. The security team needed full coverage across the OWASP Top 10, including vulnerability classes that conventional DAST tools notoriously miss: broken access control, race conditions, and business logic abuse.

They also wanted clarity on a question that comes up in nearly every AI security conversation: in an AI-powered engagement, what exactly is the AI doing, and what role does the human expert play?

THE APPROACH:

Bishop Fox deployed its Cosmos AI security testing platform against the application, followed by expert human triage and supplementary manual testing. The engagement followed a structured five-phase pipeline:

- **Phase 1: Early Reconnaissance:** Automated reconnaissance, technology stack fingerprinting, and attack surface mapping. The platform identified the application framework (Flask/Python), deployment environment (Nullstone PaaS), and confirmed no WAF was present.
- **Phase 2: Authenticated Crawling:** Authenticated crawling using admin-level credentials to fully map the application surface, discovering 20+ distinct API endpoints including user management, account operations, transfer endpoints, trading/order APIs, and administrative interfaces.
- **Phase 3: Content Distillation and Task Assignment:** Crawled content was analyzed and distilled into actionable testing tasks. AI orchestration assigned testing tasks across 8 specialized modules based on identified attack
- **Phase 4: AI Testing:** Each of the 8 modules (RECON, INTRUSION, ESCALATION, INJECTION, EXTRACTION, MANIPULATION, INFRASTRUCTURE, and CHAINS) operated as an independent AI agent with domain-specific expertise. All 8 completed successfully in 3 hours and 17 minutes.
- **Phase 5: Human Expert Triage:** A Bishop Fox security consultant reviewed all automated findings, removed duplicates, performed handson validation, calibrated severity per Bishop Fox standards, and conducted supplementary manual testing.

WHAT THE AI FOUND:

Cosmos AI generated 35 candidate findings across the 3-hour assessment. After deduplication and expert triage, 12 were confirmed as true vulnerabilities. Every finding delivered to the client was validated, resulting in a final report with zero false positives.

A particularly revealing result came from the platform's business logic testing. During its assessment of the transfer API, Cosmos AI autonomously attempted a negative dollar amount transfer, sending -$1,000,000 between accounts. The application accepted the request, immediately crediting $1M to the attacking account. No signature, rule, or scanner template produced this test case. The AI reasoned about how the transfer function should behave, hypothesized an abuse scenario, executed it, and confirmed the impact. This is the class of vulnerability that has historically separated skilled human testers from automated tools.

Across the assessment, Cosmos AI reliably surfaced findings in vulnerability classes that have historically resisted automation:

- **Broken Access Control (IDOR):** The AI identified multiple instances of broken access control, including IDOR on user data and financial endpoints. This is among the most prevalent and consequential vulnerability classes in modern web applications.
- **Business Logic: Negative Transfer Fraud:** The application allowed transfers of negative dollar amounts, enabling an attacker to effectively steal funds from other accounts. Cosmos AI demonstrated this by executing a -$1,000,000 transfer, immediately crediting $1M to the attacker account. The human expert elevated this finding from Low to High severity.

- **Business Logic: Race Condition:** Lack of transaction locking allowed five simultaneous transfer requests to execute without conflict, multiplying the source balance 5x and overdrawing the account. Both business logic findings required severity adjustment upward by the human expert, underscoring the value of expert calibration.
- **Command Injection:** Cosmos AI identified command injection on a report-generation endpoint enabling OS-level command execution, confirmed High severity after expert adjustment from Critical per Bishop Fox guidelines.
- **SSRF (Two Vectors):** Two SSRF vulnerabilities were identified: one enabling cloud metadata endpoint access (AWS-style 169.254.x.x), the other enabling internal network scanning.

WHAT THE HUMAN EXPERT ADDED:

With the automated phase complete, a Bishop Fox consultant conducted expert triage and supplementary manual testing. Because Cosmos AI had already covered the mechanical vulnerability scanning, the consultant focused exclusively on the classes of testing that require human judgment: application logic, authentication design, and code-level reasoning. The result was 8 additional findings that the AI did not surface, confirming that AI and human expertise are complementary rather than interchangeable.

- **Missing Authentication:** An unauthenticated report generation endpoint and a database reset endpoint, both exploitable without authentication.
- **Insecure Deserialization:** Insecure object deserialization, a vulnerability class requiring manual code-path reasoning that current automated tools rarely surface.
- **Sensitive Information Disclosure:** A debug endpoint exposing internal configuration and secrets.
- **SQL Injection:** SQL injection on the account search endpoint.
- **Weak Cryptography:** Weak cryptographic implementation in token handling.

The expert also eliminated 11 false positives from the AI's candidate set. Examples included the AI flagging admin-on-admin access as IDOR (expected behavior for that role), reporting JWT timing variance within normal bounds, and flagging frontend URLs rather than API authentication endpoints. This triage step is what ensures the client receives only confirmed,

THE ROLE OF SEVERITY CALIBRATION:

One of the most important human contributions in this engagement was severity calibration: the expert adjusted the severity rating on 9 of the 12 AI-confirmed findings (75%), both upward and downward, to reflect real-world exploitability and Bishop Fox's severity standards.

Severity calibration is where domain expertise and client context enter the engagement. An AI can detect that a negative transfer succeeds; a human expert understands that in a financial services environment, this represents fraud risk that warrants an immediate High severity rating. This is the core value of the human-in-the-loop model: Cosmos AI operates at machine speed and breadth, while the expert ensures every finding is graded against real-world business impact. The result is a final report containing 100% confirmed true positives with zero false positives.

CONFIRMED FINDINGS — COMPLETE INVENTORY:

All 20 confirmed findings from the engagement, with source attribution:

FINDING	SEVERITY	SOURCE
Insecure Password Reset	High	Cosmos AI
Insufficient Authorization Controls (IDOR)	High	Cosmos AI Human Expert
Missing Authentication — Reports & DB Reset	High	Human Expert
Insecure Object Deserialization	High	Human Expert
Sensitive Information Disclosure (/api/debug/info)	High	Human Expert
SQL Injection (/api/search/accounts)	High	Human Expert
Arbitrary Command Injection (/api/reports/generate)	High	Cosmos AI
SSRF (/api/documents/fetch, /api/documents/upload)	High	Cosmos AI
Insecure Input Validation (Negative Transfers)	High	Cosmos AI
Race Condition	High	Cosmos AI
Weak Cryptography	Medium	Human Expert
Debug Mode Enabled	Medium	Cosmos AI
Improper Session Management	Medium	Cosmos AI Human Expert
Cross-Site Scripting (XSS) — User Profile	Low	Cosmos AI
User Enumeration	Low	Cosmos AI
Banner / Version Information Disclosure	Low	Human Expert
Weak Password Requirements	Low	Cosmos AI
Missing Security Headers	Info	Cosmos AI
Lack of Malware Detection	Info	Human Expert
Cross-Site Scripting (XSS) — Document Retrieval	Info	Human Expert

OWASP TOP 10 COVERAGE:

Confirmed vulnerabilities were identified across 9 of 10 OWASP Top 10 categories, reflecting the breadth advantage of AI-driven testing combined with human expertise. A09: Security Logging & Monitoring Failures was not in scope for this engagement.

OWASP CATEGORY	STATUS	FINDINGS
A01: Broken Access Control	Tested - **Vulnerable**	IDOR, Missing Function-Level Access Control, Admin Bypass, Sensitive Data Exposed
A02: Cryptographic Failures	Tested - **Vulnerable**	Weak Cryptography (manual)
A03: Injection	Tested - **Vulnerable**	Command Injection, SSRF, XSS (manual/automated), SQL Injection (manual)
A04: Insecure Design	Tested - **Vulnerable**	Insecure Password Reset, Negative Transfers, Race Conditions
A05: Security Misconfiguration	Tested - **Vulnerable**	Debug Endpoints, Missing Headers, Version Disclosure
A06: Vulnerable Components	Tested - **Vulnerable**	Outdated Werkzeug 2.0.1
A07: Identification and Authentication Failures	Tested - **Vulnerable**	Lack of Token Expiration, Token Reuse
A08: Software and Data Integrity Failures	Tested - **Vulnerable**	Insecure Deserialization (manual finding)
A09: Security Logging & Monitoring Failures	Not tested (not in scope)	N/A
A10: Server-Side Request Forgery (SSRF)	Tested - **Vulnerable**	/api/documents/fetch **and** /api/documents/upload

WHAT WAS TESTED AND FOUND NOT VULNERABLE:

This section documents what was tested and found secure, not just what was found to be vulnerable. The following controls were confirmed to be properly implemented:

TEST CATEGORY	WHAT WAS TESTED	RESULT
TLS/SSL Configuration	**Protocol versions, cipher suites, certificate validity**	Properly configured (TLS 1.2+, strong ciphers)
JWT Signature Verification	**Algorithm confusion, signature stripping, key confusion**	Properly implemented (HS256)
SSTI	**Template injection via Jinja2**	Properly mitigated
Path Traversal	**Directory traversal via file operations**	No file operations exposed
DNS Zone Transfer	**AXFR requests against nameservers**	Properly denied
Default Credentials	**Common admin passwords**	No defaults found
Backup Files	**.bak, .old, .swp discovery**	None found
HTTP Verb Tampering	**Using PUT/PATCH/DELETE to bypass controls**	Properly rejected

WHY THIS MATTERS FOR YOUR SECURITY PROGRAM:

This engagement illustrates what AI-powered application penetration testing delivers in practice: not a replacement for expert testers, but an approach that produces better outcomes than either alone.

- Speed and coverage: A 3-hour, 17-minute automated phase produced 12 confirmed findings, including high-severity business logic vulnerabilities that most DAST tools miss.
- Accurate signal-to-noise: Rigorous human triage delivered a zero-false-positive final report. Your team acts on confirmed findings, not candidate alerts requiring further investigation.
- Depth: Human experts filled the gaps AI cannot cover — insecure deserialization, missing authentication on unadvertised endpoints, SQL injection requiring semantic understanding of query construction.

ABOUT BISHOP FOX

Bishop Fox the leading expert in offensive security, providing comprehensive assessment of modern environments with continuous attack surface management, red teaming, and penetration testing for applications, cloud, network, and products. We've worked with more than 25% of the Fortune 100, half of the Fortune 10, eight of the top 10 global technology companies, and all of the top global media companies to improve their

LEARN MORE AT BISHOPFOX.COM