Round 1

Survey on Explainable Security Threat Alerts

This is a 15-20 minute survey to understand how security analysts investigate cybersecurity threats and how **Explainable AI (XAI)** can enhance their workflows as well as threat analysis. The insights from this survey will help shape research on integrating XAI into security dashboards for **better threat detection, response, and mitigation.**

What is Explainable AI (XAI)? Explainable AI (XAI) refers to AI systems that provide **human-readable explanations** for their decisions, helping security teams **understand why an alert was triggered and what analysis can be performed to address the alert.**

Example: Imagine a security tool flags an email as phishing. Instead of just labeling it "Phishing," an XAI-powered system explains:

1. **Feature 1:** The email sender domain is suspicious (92% confidence).
2. **Feature 2:** The URL in the email body is linked to past phishing campaigns (87% confidence).
3. **Feature 3:** The language in the email resembles previous phishing emails (85% confidence).

This level of transparency **helps analysts validate alerts faster** and **reduce false positives.**

Your insights are crucial in shaping the development of more efficient and user-friendly security alert systems. We aim to reduce the complexity of alert handling and support security teams to make faster, more informed decisions.

We value your expert opinion and would appreciate your feedback on the following questions:

* Indicates required question

1. Email *

____

2. What is your job title? *

____

3. Name *

____

4. What is your experience in SOC operations or as a Security Analyst, in number of years? *

Mark only one oval.

○ 1-2 years
○ 2-5 years
○ 5-10 years
○ More than 10 years

Incidence Response and Workflow

5. What are first **3-5 steps** after detecting a cybersecurity alert? Please be elaborate on how you perform root cause analysis and mitigate the threat. *

6. How long does it take on an average to resolve a critical threat? *

7. How do you prioritize alerts? *

Check all that apply.

□ Severity
□ Asset Criticality
□ Past Incidents
□ Threat Intelligence
□ Manual Review
□ Other: ____________________

Security Tools and Workflows

8. What **security tools** do you currently use? *

Check all that apply.

□ SIEM
□ XDR
□ EDR
□ IDS/IPS
□ SOAR

□ Other: ____________________

9. What is your **biggest frustration** with your current security tools? *

____

Explainable AI in Security

10. Have you used **any XAI-powered tools** in your security investigations? *

Mark only one oval.

○ Yes
○ No

11. If yes, which **XAI tools** have you used? (if no say n/a) *

____

12. If no, what kind of **explainability features** would be helpful? *

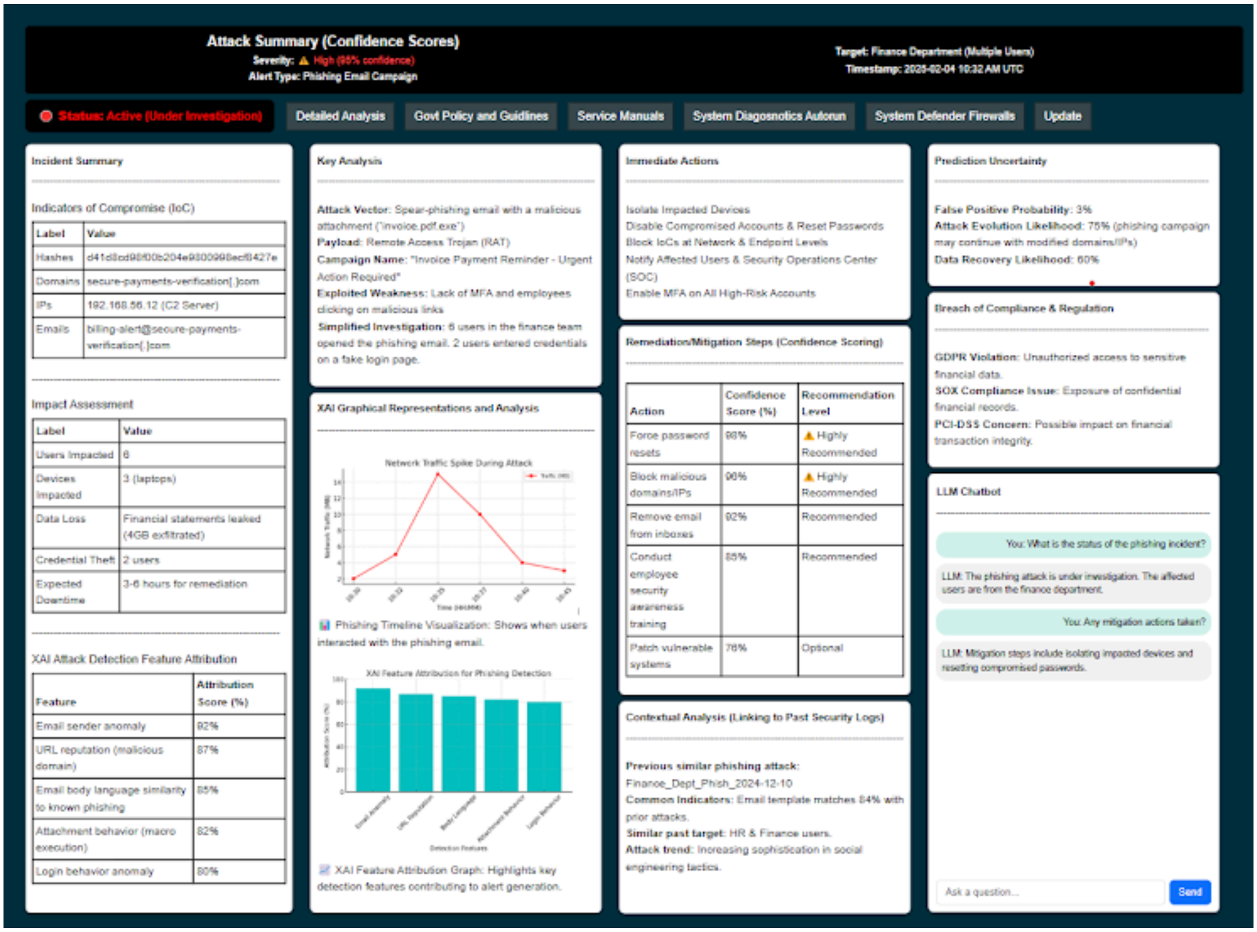
Check all that apply.

□ Confidence Scores
□ Attack Attribution
□ Alternative Explanations
□ Feature Contribution
□ Other

Feedback on XAI enabled Dashboards

Below are 2 dashboards we've designed for you to envision how XAI can be integrated into a SOC workflow. For example, the XAI Graph Representation and analysis shows important features when phishing is recognized on a client machine.

13. See a SOC dashboard 1 below that we've designed to integrate XAI into your workflows. Which explanations do you prefer and why? *



Attack Summary (Confidence Scores)
Severity: ⚠ High (95% confidence)
Alert Type: Phishing Email Campaign

Target: Finance Department (Multiple Users)
Timestamp: 2025-02-04 10:32 AM UTC

● Status: Active (Under Investigation)
Detailed Analysis
Govt Policy and Guidlines
Service Manuals
System Diagosnotics Autorun
System Defender Firewalls
Update

Incident Summary

Indicators of Compromise (IoC)

Label	Value
Hashes	d41d8cd98f00b204e9800998ecf8427e
Domains	secure-payments-verification[.]com
IPs	192.168.56.12 (C2 Server)
Emails	billing-alert@secure-payments-verification[.]com

Impact Assessment

Label	Value
Users Impacted	6
Devices Impacted	3 (laptops)
Data Loss	Financial statements leaked (4GB exfiltrated)
Credential Theft	2 users
Expected Downtime	3-6 hours for remediation

XAI Attack Detection Feature Attribution

Feature	Attribution Score (%)
Email sender anomaly	92%
URL reputation (malicious domain)	87%
Email body language similarity to known phishing	85%
Attachment behavior (macro execution)	82%
Login behavior anomaly	80%

Key Analysis

Attack Vector: Spear-phishing email with a malicious attachment ("invoice.pdf.exe")
Payload: Remote Access Trojan (RAT)
Campaign Name: "Invoice Payment Reminder - Urgent Action Required"
Exploited Weakness: Lack of MFA and employees clicking on malicious links
Simplified Investigation: 6 users in the finance team opened the phishing email. 2 users entered credentials on a fake login page.

XAI Graphical Representations and Analysis

📊 Phishing Timeline Visualization: Shows when users interacted with the phishing email.

📈 XAI Feature Attribution Graph: Highlights key detection features contributing to alert generation.

Immediate Actions

Isolate Impacted Devices
Disable Compromised Accounts & Reset Passwords
Block IoCs at Network & Endpoint Levels
Notify Affected Users & Security Operations Center (SOC)
Enable MFA on All High-Risk Accounts

Remediation/Mitigation Steps (Confidence Scoring)

Action	Confidence Score (%)	Recommendation Level
Force password resets	98%	⚠ Highly Recommended
Block malicious domains/IPs	96%	⚠ Highly Recommended
Remove email from inboxes	92%	Recommended
Conduct employee security awareness training	85%	Recommended
Patch vulnerable systems	78%	Optional

Contextual Analysis (Linking to Past Security Logs)

Previous similar phishing attack: Finance_Dept_Phish_2024-12-10
Common Indicators: Email template matches 84% with prior attacks.
Similar past target: HR & Finance users.
Attack trend: Increasing sophistication in social engineering tactics.

Prediction Uncertainty

False Positive Probability: 3%
Attack Evolution Likelihood: 75% (phishing campaign may continue with modified domains/IPs)
Data Recovery Likelihood: 60%

Breach of Compliance & Regulation

GDPR Violation: Unauthorized access to sensitive financial data.
SOX Compliance Issue: Exposure of confidential financial records.
PCI-DSS Concern: Possible impact on financial transaction integrity.

LLM Chatbot

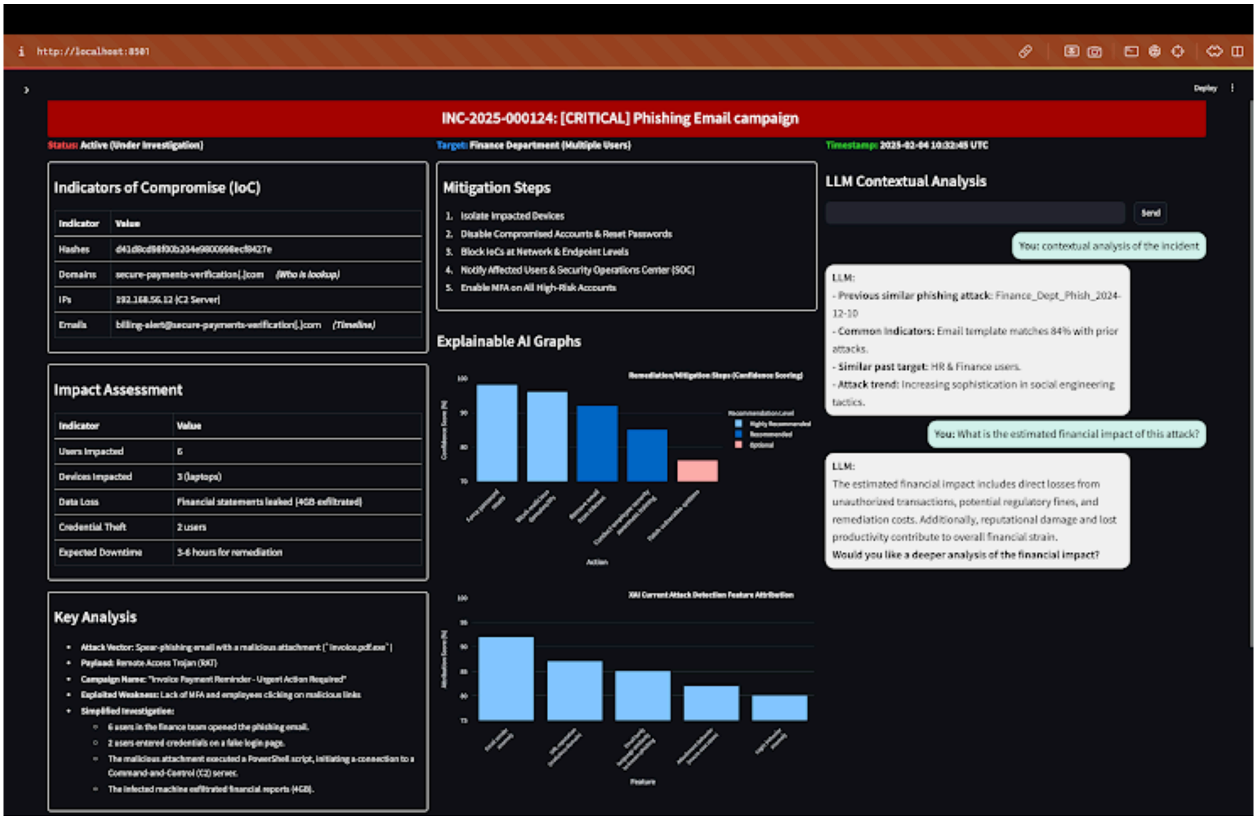
You: What is the status of the phishing incident?

LLM: The phishing attack is under investigation. The affected users are from the finance department.

You: Any mitigation actions taken?

LLM: Mitigation steps include isolating impacted devices and resetting compromised passwords.

14. See a SOC dashboard -2 below. Which explanations do you prefer and why? *



INC-2025-000124: [CRITICAL] Phishing Email campaign

Status: Active (Under Investigation)

Target: Finance Department (Multiple Users)

Timestamp: 2025-02-04 10:32:45 UTC

Indicators of Compromise (IoC)

Indicator	Value
Hashes	d41d8cd98f00b204e9800998ecf8427e
Domains	secure-payments-verification[.]com *(Who is lookup)*
IPs	192.168.56.12 (C2 Server)
Emails	billing-alert@secure-payments-verification[.]com *(Timeline)*

Impact Assessment

Indicator	Value
Users Impacted	6
Devices Impacted	3 (laptops)
Data Loss	Financial statements leaked (4GB exfiltrated)
Credential Theft	2 users
Expected Downtime	3-6 hours for remediation

Key Analysis

- **Attack Vector:** Spear-phishing email with a malicious attachment (`Invoice.pdf.exe`)
- **Payload:** Remote Access Trojan (RAT)
- **Campaign Name:** "Invoice Payment Reminder - Urgent Action Required"
- **Exploited Weakness:** Lack of MFA and employees clicking on malicious links
- **Simplified Investigation:**
  - 6 users in the finance team opened the phishing email.
  - 2 users entered credentials on a fake login page.
  - The malicious attachment executed a PowerShell script, initiating a connection to a Command-and-Control (C2) server.
  - The infected machine exfiltrated financial reports (4GB).

Mitigation Steps

1. Isolate Impacted Devices
2. Disable Compromised Accounts & Reset Passwords
3. Block IoCs at Network & Endpoint Levels
4. Notify Affected Users & Security Operations Center (SOC)
5. Enable MFA on All High-Risk Accounts

Explainable AI Graphs

LLM Contextual Analysis

You: contextual analysis of the incident

LLM:
- **Previous similar phishing attack:** Finance_Dept_Phish_2024-12-10
- **Common Indicators:** Email template matches 84% with prior attacks.
- **Similar past target:** HR & Finance users.
- **Attack trend:** Increasing sophistication in social engineering tactics.

You: What is the estimated financial impact of this attack?

LLM:
The estimated financial impact includes direct losses from unauthorized transactions, potential regulatory fines, and remediation costs. Additionally, reputational damage and lost productivity contribute to overall financial strain.
Would you like a deeper analysis of the financial impact?

Insights on XAI

15. What challenges do you face when **understanding AI-driven security alerts**? *

16. What are your biggest **concerns** about using AI in cybersecurity? *

17. Will you be open to have a short 30 minute follow-up call with our research team? *