Báo cáo nội dung các bài trình bày – Security Bootcamp 2026

Nexus Lab, VNU Information Technology Institute

> ***Security Bootcamp 2026 – Agentic Security*** *· 10–12/09/2026 · KS Hai Bà Trưng, TP. Buôn Ma Thuột, Đắk Lắk Tổ chức: Hiệp hội Internet Việt Nam (VIA) · Agenda gốc: https://securitybootcamp.org.vn/agenda*
>
> *Báo cáo tổng hợp nội dung* ***26/27 bài trình bày có tài liệu công khai*** *(tính đến 15/09/2026; tham luận #11, Apple Secure Kernel, là nghiên cứu đang thực hiện nên không công bố slide), được* ***tổ chức lại theo 6 chủ đề****. 5 bài có bản cập nhật v1.1 từ ban tổ chức; nội dung báo cáo cập nhật theo bản mới nhất và ghi nhận khác biệt tại mục "Nhật ký cập nhật". Số liệu, CVE, link trong báo cáo được trích nguyên trạng từ slide.*

Mục lục theo chủ đề

Tổng hợp ngắn gọn theo chủ đề

Mỗi bài có tóm tắt 3 câu (nội dung / vấn đề giải quyết / kết quả). **Bấm vào tiêu đề để đọc phần tóm tắt chi tiết tương ứng.**

Chủ đề 1 – Bối cảnh & chiến lược phòng thủ trong kỷ nguyên AI

#1. **AI Nguồn Mở và mối quan hệ giữa AI tạo sinh và sở hữu trí tuệ** *TS. Lê Trung Nghĩa (InOER, AVU&C)*

Bài trình bày về khái niệm AI Nguồn Mở theo định nghĩa OSI (4 tự do: Sử dụng, Nghiên cứu, Sửa đổi, Chia sẻ) và mối quan hệ giữa AI tạo sinh với sở hữu trí tuệ dưới góc nhìn pháp lý Việt Nam. Vấn đề được đặt ra là rủi ro vi phạm bản quyền và SHTT khi ứng dụng AI, trong bối cảnh Luật Trí tuệ nhân tạo 134/2005/QH15, Bộ luật Hình sự Điều 225 và Công điện 38/CĐ-TTg siết chặt chế tài, cùng tình trạng Việt Nam nằm Top 10 quốc gia dùng phần mềm lậu. Kết luận đưa ra là nên ứng dụng và phát triển AI Nguồn Mở (kèm khung MOF của Linux Foundation và chỉ số minh bạch dữ liệu AIDTI) để hệ AI tuân thủ luật và tránh 'xây nhà trên móng của người khác'.

#9. Từ thách thức An ninh mạng đến nghiên cứu và đổi mới *Đỗ Đình Huân (VinSOC)*

Bài chia sẻ cách một đơn vị security mới thành lập bảo vệ một hệ sinh thái đã vận hành nhiều thập kỷ (42.000+ thiết bị, 900 hệ thống, 6 quốc gia) khi tốc độ khai thác lỗ hổng đã chuyển từ 'ngày' sang 'giờ'. Vấn đề cốt lõi là cuộc chiến phi đối xứng: chi phí tấn công tiến về gần 0 nhờ AI, trong khi chi phí phòng thủ không giảm, khiến chiến lược 'mua thêm, triển khai thêm' chạm trần. Kết quả là mô hình 3 đòn bẩy đã kiểm chứng: threat modeling ở giai đoạn thiết kế (giảm 40% lỗ hổng lặp lại), 312 Security Champions tự phát hiện 2.000+ lỗ hổng không tăng biên chế, và nhân bản chuyên gia qua kho tri thức + Red Team đối kháng khép kín vòng cải tiến.

Chủ đề 2 – Bảo mật AI & Agent: tấn công và kiểm soát

#2. MIND INCEPTION – Ai Đang Điều Khiển AI? *Nguyễn Thúy Hằng, Nguyễn Đức Kiên (VCS AILAB × AI4LIFE)*

Bài trình bày nghiên cứu red-teaming khiến mô hình AI tự kết luận một webshell là file sạch (CLEAN) thông qua kỹ thuật 'tự sinh ngữ cảnh đánh lừa chính mô hình' trên OpenAI gpt-oss-20b. Vấn đề được nghiên cứu là khả năng điều khiển gián tiếp hành vi LLM không qua prompt trực tiếp, dựa trên giả thuyết 'một chuỗi đã từng được sinh ra là một chuỗi đã được tối ưu xác suất'. Kết quả: kỹ thuật MIND INCEPTION đạt ASR 49,6–100% trên 10 mô hình open-weight (vượt xa Raw prompt ≈0% và CoT), trace còn chuyển giao chéo giữa các mô hình (84–100%), cho thấy không gian xác suất của các LLM hiện nay đã hội tụ.

#3. When AI can act, Who controls the AI? *Nguyễn Tuấn Hưng (VNPT Cyber Immunity)*

Bài trình bày khung AI Security Control Layer kiểm soát xuyên suốt 5 tầng DATA, PROMPT, MODEL, AGENT, TOOL với 4 nguyên tắc DETECT, DECIDE, ENFORCE, AUDIT. Vấn đề là firewall/IAM/DLP truyền thống không 'hiểu' luồng prompt, context, agent và tool/MCP, trong khi pháp lý dữ liệu cá nhân (Luật 91/2025/QH15, Nghị định 356/2025, phạt tới 5% doanh thu) bắt buộc kiểm soát khi AI đã hành động thay con người. Kết quả là kiến trúc AI Guardrails hoàn chỉnh: DLP check đầu vào, policy theo người dùng/phòng ban, tool governance với kill-switch và audit trail cho từng hành động của AI Agent.

#18. Buffer Overflow trong AI – Tại sao Filter không thể cứu chúng ta? *(v1.1 đổi tên: "Prompt Injection Bootcamp") Đinh Công Thái, Nguyễn Sơn (Viettel Cyber Security)*

Bài đối chiếu bài học memory safety (stack canary, NX/W^X, Rust) với bài toán prompt injection và đề xuất kiến trúc mô hình mới chống prompt extraction. Vấn đề là các filter/detector (kể cả trước GCG và NeuralExec, hai kỹ thuật tấn công tối ưu bằng gradient) chỉ 'phát hiện' chứ không bảo đảm về mặt kiến trúc rằng data độc không thể biến thành control. Kết quả là đề xuất IMA (Instruction Memory Adapter): tách instruction channel khỏi token sequence, decoder cross-attention vào Instruction Memory, với lập luận I*

không thể bị invert từ attention output nên prompt extraction về mặt kiến trúc là không thể.

#23. From Lab to Prod: MLOps Strategies for Securing Agentic AI Deployments *Nguyễn Thu Thuỷ (Sophie), Chitta Group*

Bài trình bày chiến lược đưa agentic AI từ lab lên production an toàn, mở đầu bằng sự cố thật: agent đọc PDF rồi tự gửi nội dung nội bộ ra ngoài, không có CVE, không lỗ hổng code. Vấn đề là agent mang State + Loop + quyền hạn khiến mỗi nội dung nó đọc đều có thể trở thành chỉ thị (data becomes instruction), permission creep và trust chain qua MCP bên thứ ba, trong khi metric hạ tầng (latency/error) vẫn xanh. Kết quả là khung 4 lớp bắt buộc (bounded delegation, tool boundary 3 mức với two-man rule, context isolation + egress check, audit trail có context_hash), cách đo hành vi bằng drift detection, và hướng nghiên cứu stability gate đo khoảng cách hành vi D so với phiên bản đã duyệt bảo mật.

Chủ đề 3 – Agentic AI vận hành SOC & phát hiện

#4. Autonomous SOC: From Model Capabilities to Verified Autonomy *Đặng Trần Thái (VinSOC)*

Bài trình bày phương pháp xây dựng Autonomous SOC với AI agent có situational awareness, reasoning, planning và khả năng hành động bảo mật trong vòng lặp khép kín Perceiving, Reasoning, Planning, Acting, Verifying, Learning. Vấn đề là các rào cản khi triển khai: hallucination, prompt injection, AI quyết sai nhưng vẫn có quyền hành động, tool abuse và thiếu cơ chế đánh giá/xác minh hành động. Kết quả là kiến trúc Cyber Intelligence Hub (chuẩn hóa schema, entity relationship, data mining AI/ML với trung tâm AI thống nhất) kèm bộ metrics chất lượng, độ tin cậy làm baseline xác lập mức tự chủ của agent.

#7. Agentic Cloud Detection & Response: From Alert to Action *Phạm Thế Thiện (VSEC)*

Bài trình bày tích hợp Agentic AI vào quy trình Cloud Detection & Response của VSEC, từ cảnh báo đến hành động ứng phó trên môi trường multi-cloud. Vấn đề là SOC cloud truyền thống quá tải (alert fatigue, log phân tán trên CloudTrail/VPC Flow/K8s/SaaS, alert thiếu ngữ cảnh, MTTD/MTTR chậm) trong khi sự cố cloud tăng 60% và supply-chain attack tăng gấp đôi (Wiz H1 2026). Kết quả là vòng lặp Agentic Threat Intelligence hoạt động thực tế: agent tự enrichment, correlate, trả verdict kèm confidence và recommendation cho từng alert, giúp giảm alert fatigue và tăng tốc MTTD/MTTR với human-in-the-loop ở quyết định block/escalate/contain.

#10. Threat Hunting Lessons from the Field: A Practical Approach for Shifting from Triage to Disruption *James Murphy (Trellix)*

Bài chia sẻ kinh nghiệm threat hunting thực chiến từ Field CTO Trellix APJ, chuyển trọng tâm từ triage alert sang disruption adversary. Vấn đề là hunting đo bằng số alert/query đã lỗi thời; machines và AI xử lý dữ liệu cực nhanh nhưng vẫn cần con người quyết định hunt cái gì, vì sao tín hiệu quan trọng và điều tra đi tiếp đâu. Kết quả là phương pháp hunting dựa trên nhiều tín hiệu kết hợp thay vì IOC đơn lẻ, minh chứng qua 2 case study (campaign JSCeal nhắm Đông Nam Á, phishing FIFA World Cup của Bitter APT) cùng dịch vụ Trellix SecondSight.

#24. SOC in the age of AI Agent *Bùi Đức Trường (Misa)*

Bài trình bày kiến trúc Agentic SOC đầy đủ 4 tầng (Knowledge Engine, Agent Layer, Human Oversight, Tools & Integrations) và nguyên tắc 'trust before autonomy'. Vấn đề là SOC truyền thống bị bào mòn bởi dữ liệu (200 triệu log/ngày, 200 alert/ngày, 25–30 phút/alert) trong khi attacker breakout chỉ 27 giây, và harness chỉ giảm rủi ro chứ không loại bỏ. Kết quả nổi bật là use case Alert Triage closed-loop giảm 25–30 phút xuống ~2 phút (~93% nhanh hơn) với HITL, và triết lý governed autonomy: autonomy mở rộng/thu hẹp theo bằng chứng, 'an Agentic SOC is a system, not just a model'.

Chủ đề 4 – AI Agent trong công tác phân tích & khai thác lỗ hổng

#5. Combining Static Analysis, Symbolic Execution, and LLM for Vulnerability Scanning *Ngô Trần Quang Duy (CMC Telecom)*

Bài trình bày cách kết hợp 3 kỹ thuật quét lỗ hổng (static analysis, symbolic execution, fuzzing) với LLM điều phối thành phương pháp SAILOR. Vấn đề là mỗi kỹ thuật đều có điểm yếu riêng: static analysis false positive cao (412 findings/target), symbolic execution cần harness viết tay và path explosion, fuzzing không chạm deep path. Kết quả là pipeline SAILOR (Static Analysis Informed and LLM-Orchestrated Symbolic Execution): static analysis chỉ ứng viên, LLM xây harness, symbolic execution chứng minh đường tới sink, kèm demo và đánh giá thẳng thắn về giới hạn (vẫn phụ thuộc CodeQL, ASan chưa khẳng định exploitability).

#16. HUNTing 0-day In AI Era *Lê Thế Thắng (LilThawg29), One Mount Group*

Bài kể hành trình dùng AI assistant săn 0-day cho Pwn2Own Berlin 2026 (đội Việt Nam đầu tiên) trên target LiteLLM, AI Gateway mà 100% LLM traffic đi qua. Vấn đề là chọn target và phương pháp săn bug khi thời gian/nhân lực có hạn, cộng với rủi ro AI 'reward hacking' (agent lách shortcut, chốt trong 1–2,5 lượt mà không đi đúng lỗ hổng). Kết quả là 2 chuỗi RCE unauthenticated hoàn chỉnh: SQLi → Postgres superuser → file write → admin token seeding → sandbox escape → RCE root (~30 giây), và Host Header Confusion → control plane → privileged API key → MCP stdio → RCE root, cùng phương pháp kiểm chứng AI findings 4 lớp.

#20. Ứng dụng AI Agent trong phân tích mã độc – Skill, IOC, Deobfuscate *Tạ Đăng Vinh, Nguyễn Duy Bình (VNPT)*

Bài trình bày việc chuẩn hóa phân tích mã độc bằng AI Agent thông qua bộ Skill (file Markdown quy định quy trình) và MCP kết nối IDA Pro. Vấn đề là prompt tự do khiến agent bỏ bước, chọn sai tool và kết quả khó kiểm chứng, trong khi giai đoạn phân tích ban đầu (hash, strings, imports, IOC) lặp lại và tốn thời gian. Kết quả là bộ skill hoạt động với 5 playbook (pe_dotnet, apk, office_script, web_payload, unix_binary) cho output chuẩn (AGENTS.md, REPORT.md, evidence, IOC) được thử nghiệm trên mẫu lừa đảo thật bằng Codex gpt-5.6-sol, với khuyến nghị bắt buộc bước xác minh lại do chất lượng phụ thuộc năng lực mô hình.

#22. Breaking Access Control with Fuzzer: Hunting for BAC Zero-Days *Vũ Chí Thành, Hoàng Xuân Quân, Đinh Thái Sơn (VinSOC)*

Bài trình bày phương pháp phát hiện Broken Access Control (BAC, OWASP risk #1) bằng cách "ask the sink, not the response": theo dõi điểm kiểm tra quyền trong runtime qua debugger thay vì so sánh response giữa 2 phiên. Vấn đề là so sánh response truyền thống cho ra false positive cả hai chiều, và nhiều endpoint (như CVE-2026-49447 Cosmos Server) không mang id nào để so sánh. Kết quả là phương pháp 5 stages (crawl, profile, replay,

observe at sink qua JDWP/Delve, verdict) tìm ra CVE thật: UniFi OS CRITICAL 9.9 (chưa vá) và CVE-2026-63479 (DevGuard), kiểm chứng bằng RUN 1/RUN 2 trên build có bug và build đã vá.

#26. Read, Grep, Glob & Graph: Xây dựng AI Agent cho Code Review *Quốc Hùng, Nguyễn Thanh Tùng (FPT Smart Cloud)*

Bài trình bày kiến trúc AI agent code review tối giản chỉ với 4 tools (Read, Grep, Glob, Graph) vận hành production trên Kafka tại FPT Smart Cloud. Vấn đề là SAST sinh quá nhiều finding mà findings không đồng nghĩa vulnerabilities, và grep/read thuần không trả lời được 'cái gì đi tới cái gì' trên repository lớn (symbol trùng tên, hàng trăm nơi xuất hiện). Kết quả là pipeline triage finding tự động (scanner đưa giả thuyết, agent luận giải reachability/dataflow) với step/tool budget thay token budget, demo case Snipe-IT CVE-2025-63601 (agent tự kiểm chứng ZIP restore ghi PHP vào public/uploads qua continue 2 bỏ qua extension guard) và các kinh nghiệm vận hành model opensource (corrective feedback, chống doom-loop).

Chủ đề 5 – Nghiên cứu tấn công chuyên sâu

#6. Navigating A New Cyber Reality – Bring Your Own EDR: How to turn EDR into a Trojan Horse *Lê Gia Quý (Akamai)*

Bài trình bày kỹ thuật BYOEDR: biến chính agent EDR được bảo vệ bởi Windows PPL thành cửa xâm nhập, trên bối cảnh bùng nổ API/AI bot (73% doanh nghiệp không biết API nào trả dữ liệu nhạy cảm). Vấn đề là PPL và các COM object của EDR (SentinelHelper của SentinelOne) thiếu validation caller, cho phép local admin dump tiến trình PPL và trích COM secrets để inject vào Windows Defender. Kết quả là chuỗi kỹ thuật được chứng minh end-to-end (exposed COM object → PPLSystem dump → COM secrets extraction → unsigned DLL indicator.dll chạy trong MsMpEng.exe, khắc phục 5 trở kiện kỹ thuật), kèm khuyến nghị hardening compile static và dữ liệu từ Glasswing/OpenAI TAC.

#12. Securing Connected Vehicles: From Automotive Pentesting to AI-Driven Security *Nguyễn Đình Mạnh, Nguyễn Minh Ngọc (Vietsol)*

Bài trình bày thực trạng automotive security (xe 100+ ECU, 4 hướng tấn công: backend, wireless, in-vehicle, MCU) và phương pháp pentest ô tô của Vietsol. Vấn đề là bề mặt tấn công xe mở rộng qua autonomous driving, V2X, OTA, EV payment trong khi các flow 'silent UX' (endpoint ẩn không có UI) không bao giờ được pentest UI-based chạm tới. Kết quả là 4 case study khai thác thật (BOLA qua endpoint refresh_session ẩn → account takeover CWE-306 CRITICAL, UDS SecurityAccess seed 0xAA55 tĩnh → brute force 2 byte, RF replay keyless, HID injection treo màn hình) và nền tảng CANPORT tự động hóa TARA/pentest đạt ISO/SAE 21434, giảm 20% effort, nhanh gấp 3 lần orchestration.

#14. Tấn công chuỗi cung ứng: Khi mối đe dọa xuất phát từ nguồn tin cậy nhất *Đoàn Minh Long (VCSTI)*

Bài trình bày bức tranh supply chain attack 2025–2026 chuyển dịch sang AI agent & OIDC (model, skill, MCP, trusted publishing vào chính trust boundary). Vấn đề là attacker xâm phạm trust boundary trước (token CI không revoke đủ, slopsquatting tên package AI 'bịa' ra, SKILL.md độc hại, gói npm typosquat) để automation phân phối impact sau. Kết quả là chuỗi thời gian sự cố TeamPCP 3/2026 (Trivy → LiteLLM v1.82.7/1.82.8 trên PyPI, stealer SANDCLOCK, 2.500+ tổ chức, 153 GB credentials), 3 CVE AI supply chain (ChromaDB

CVE-2026-45829 pre-auth RCE, HF Transformers CVE-2026-4372/CVE-2026-5241), sự cố postmark-mcp BCC ẩn, và khuyến nghị 'treat SKILL.md as code'.

#15. When Async Programming Becomes A Real Risk: Weaponizing Coroutine To Defeat Control-Flow Integrity *Hoàng Tiến Minh (VinSOC)*

Bài giới thiệu CFOP (Coroutine Frame Oriented Programming), kỹ thuật hijack control-flow mới dựa trên coroutine frame của C++20 nằm trên heap. Vấn đề là Intel CET (Shadow Stack + Indirect Branch Tracking) tuy tiêu diệt ROP/JOP/COP nhưng không bảo vệ coroutine frame, và mọi nỗ lực backward-compatibility với CFI (như coroutine) mở bề mặt tấn công mới. Kết quả là CFOP đánh bại hardware-enforced CFI trên Windows/MSVC và Linux x64/GCC (ARM64e PAC vẫn chặn được), cùng framework cfopwn tự động hóa end-to-end exploit generation dùng LLM prune control-flow trước symbolic execution.

#21. ICMPWN – From PingPong to PingPWN, Using ICMP as a Covert C2 Channel *Nguyễn Quang Minh, Phạm Tiến Mạnh (CyPeace)*

Bài trình bày icmpwn, C2 framework tự xây chạy hoàn toàn trong ICMP khi mọi outbound TCP/UDP/DNS/HTTP bị chặn trong một red team engagement thật. Vấn đề là các tool sẵn có không đáp ứng: icmpsh/Invoke-PowerShellIcmp chỉ là reverse shell không mã hóa và dễ bị log, CS-EXTC2 cần Cobalt Strike 3.500 USD/năm, còn Havoc/Sliver/Mythic không có ICMP transport. Kết quả là framework full-featured: 9-byte header + AES-256-GCM, multi-session (4-byte SESSION_ID), fragmentation theo RFC 792 với payload 1.472 B, EDR bypass bằng indirect syscall (ntdll từ đĩa), AMSI patch, ETW silenced, kèm post-ex BOF, in-memory .NET, nanodump, SOCKS5 over ICMP.

#25. Vibe Hacking: How AI Agent Gained Access to Thousands of TP-Link Tapo Cameras Worldwide *Đỗ Nhật Thái, Trần Lê Đăng Khôi (/vibe/hack)*

Bài kể cách dùng AI agent (Codex + IDA Pro MCP) reverse engineering firmware TP-Link Tapo C200 để tìm và khai thác lỗ hổng, từ firmware extraction (kèm decryption bằng tool tp-link-decrypt) tới RCE chain. Vấn đề là firmware IoT MIPS (Ingenic T23, Linux 3.10.14, BusyBox) chứa nhiều attack surface không ai rà, và prompting mơ hồ ('find all the bugs') không hiệu quả bằng prompt có workspace, setup, goal và môi trường test rõ ràng. Kết quả là 6 lỗ hổng báo TP-Link trong 2 ngày (15–16/04/2026), trong đó chuỗi unauthenticated RCE hoàn chỉnh: auth bypass qua cnonce rỗng → enable TMPD daemon TCP/20002 → command injection qua MacTool tftp → reverse shell qua TFTP; TP-Link vá ngày 10/07/2026 và yêu cầu 30 ngày trước công bố.

Chủ đề 6 – Mã độc, hạ tầng C2 & điều tra

#8. Phân tích hệ sinh thái gian lận chuyển khoản giả mạo tại Việt Nam *Bùi Trọng Hiếu (Viettel Threat Intelligence)*

Bài mổ xẻ một nạn nhân thật (chủ quán ăn thấy 'Giao dịch thành công' giả) để truy ngược toàn bộ hệ sinh thái gian lận chuyển khoản giả mạo tại Việt Nam. Vấn đề là mọi kỹ thuật giả bill/app đều khai thác điểm mù giữa Presentation Layer và Data Layer, trong khi mô hình Fraud-as-a-Service (dựng web 2 triệu VND, thuê tài khoản giả 60K/giờ – 25 triệu) hạ thấp rào cản kỹ thuật cho hàng trăm đối tượng. Kết quả là reverse engineering hạ tầng C2 trong strings.xml của APK (C2 HTTP, WebSocket 4321, Firebase, Telegram Bot, x.chungchi.one), link-analysis 15 mối liên hệ (email miltonfar****@gmail.com đăng ký 16 tên miền), cơ chế app giả truy vấn API tra cứu tên chủ tài khoản hợp pháp (endpoint /api/app/find_bank), và khuyến nghị 4 nhóm kèm hướng truy vết dòng tiền doanh thu FaaS.

#13. MustangPanda and PlugX: State-Sponsored APT or Cyber Mercenary? *Trần Trung Kiên (m4n0w4r)*

Bài phân tích nhóm APT MustangPanda và mã độc PlugX từ góc nhìn 10 năm lịch sử, đặt câu hỏi đây là chủ nhà nước hay 'lính đánh thuê mạng'. Vấn đề là hạ tầng và kỹ thuật của nhóm tái sử dụng qua nhiều chiến dịch (LNK loader, shellcode, DLL sideloading) khiến IOC cũ nhanh lỗi thời, đòi hỏi quy trình triage, deep dive và extractor tự động hóa có kiểm chứng trên nhiều mẫu. Kết quả là bộ script trích xuất config PlugX hoạt động trên nhiều mẫu mới (kể cả chuỗi lây qua multi-stage SVG), cùng nhận định 'APT operators do not need to remain invisible forever, they only need to remain undetected long enough to achieve their objectives'.

#17. EtherHiding và EtherRAT: Lạm dụng Blockchain để phân giải C2 và duy trì hạ tầng mã độc *Trần Trọng Huy (BlueCyber)*

Bài phân tích kỹ thuật EtherHiding: malware lưu cấu hình C2 trong smart contract Ethereum/Polygon và đọc qua JSON-RPC eth_call không cần ví, không gas. Vấn đề là IOC tĩnh nhanh hết hiệu lực khi C2 luân chuyển on-chain (majority vote trên 7–14 RPC song song), dữ liệu on-chain không thể takedown, và 4 chiến dịch EtherRAT thực tế (Unit 42, DFIR Report, Atos, Threat Labs/ClickFix) nhắm từ quản trị viên DevOps tới người dùng phổ thông. Kết quả là bộ khuyến nghị phát hiện theo hành vi (cài Node.js bất thường, chuỗi cmd → node.exe → cmd/powershell, POST eth_call lặp kỳ đều 30s/5 phút), đánh giá đúng mức chặn contract address (IOC tốt để hunting nhưng không phải điểm chặn mạng), và kỹ thuật NullReceiver giấu IP C2 trong trường to của giao dịch 0 ETH.

#19. OSINT hay stalk? Woman! *Huỳnh Ngọc Khánh Minh (Chống Lừa Đảo)*

Bài kể case OSINT điều tra tổ chức tội phạm (CASE-ORG-A) chỉ bằng nguồn công khai, qua 13 bước trong 4 chặng từ LinkedIn/GitHub tới vòng người thân. Vấn đề là founder khóa tài khoản rất kỹ trong khi bề mặt công khai của tổ chức nằm ở người thân nữ (mẹ/vợ đối tượng đăng ảnh, tag, số nhà, email kinh doanh), cùng thách thức quy mô khi hàng chục tài khoản cần rà thủ công. Kết quả là xác định được vòng nhân sự mở rộng, khớp nhiều tài khoản ngân hàng với họ hàng hai bên, và định vị địa chỉ một đối tượng cầm đầu bằng đối chiếu dữ liệu rò rỉ đã phát tán, không cần công cụ trả phí hay quyền truy cập đặc biệt.

#27. When AI Agents Become Scammers – Mô phỏng đối kháng đa tác nhân để phòng thủ vishing tiếng Việt *Nguyễn Quốc Phú, Vũ Tiến Minh (VCS AILAB × AI4LIFE)*

Bài trình bày ViScamDial: thay vì sinh hội thoại lừa đảo thuần, xây chính các scammer agent đối kháng với listener agent để tạo dữ liệu huấn luyện detector vishing tiếng Việt. Vấn đề là data paradox: cuộc gọi thật chứa PII không dùng được, synthetic thông thường không đa dạng, dữ liệu dịch từ tiếng Trung/Anh không sát tập quán Việt Nam, và detector dễ báo nhầm cuộc gọi xác minh hợp lệ của ngân hàng. Kết quả là ViVOD (270M tham số, streaming detection nhân quả) bắt 98,9–100% cuộc gọi lừa đảo với 5,5% báo nhầm (đối thủ GPT/Claude/Gemini: 42,7–97,3% báo nhầm theo biểu đồ), và kết hợp 417 cuộc gọi thật + 12.886 hội thoại ViScamDial đạt 94,3% accuracy, 95,5% detection.

Quy tắc cập nhật: khi một bài có nhiều phiên bản, báo cáo **theo bản mới nhất** và ghi nhận khác biệt tại đây; bản cũ vẫn được lưu nguyên trong `Presentations/` để đối chiếu.

Phụ lục – Phương thức lập báo cáo & đảm bảo tính khách quan

Quy trình tạo lập

Báo cáo được tạo lập theo quy trình **tự động hóa có kiểm soát** trên bộ tài liệu gốc, gồm các bước sau (mọi bước đều chạy trong môi trường container cô lập, tái lập được):

1. **Thu thập dữ liệu nguồn.** Toàn bộ 25 file slide công khai (PDF/PPTX) được tải tự động từ Google Drive chính thức của ban tổ chức bằng công cụ mã nguồn mở `gdown`, nguyên trạng không chỉnh sửa, lưu tại `Presentations/`.
2. **Trích xuất nội dung máy đọc.** Mỗi file được trích xuất văn bản bằng thư viện PyMuPDF (PDF) và python-pptx (PPTX) – hai công cụ đọc theo cấu trúc file, không qua diễn giải của con người. Kết quả lưu tại `extracted/` với đánh dấu trang/slide (`===== PAGE n =====`, `===== SLIDE n =====`) cùng phần ghi chú diễn giả (`[NOTES]`) để mọi nội dung đều **truy vết ngược được về đúng trang gốc**.
3. **Bóc tách hình tổng quan.** Các slide chứa sơ đồ kiến trúc, chuỗi tấn công, agenda đồ họa được render tự động ở độ phân giải cố định (110 dpi) thành `assets/figures/`, đặt tên theo quy ước `fig-<số bài>-<mô tả>.png`.
4. **Tổng hợp theo chủ đề.** Nội dung 25 bài được phân nhóm vào 6 chủ đề dựa trên **chủ đề tự khai của từng bài** (agenda chính thức) thay vì đánh giá chủ quan của người viết. Trình tự trong từng chủ đề theo số tham luận.
5. **Trích dẫn nguyên trạng.** Số liệu định lượng (ASR, detection rate, số CVE, số tiền...), tên CVE và liên kết nguồn được chép từ slide kèm tên nguồn mà slide tự ghi (CrowdStrike, Wiz, Unit 42...), không làm tròn, không suy diễn ngoài phạm vi slide.

Cơ chế đảm bảo tính khách quan

Việc khách quan hóa được thực hiện chủ yếu bằng **xử lý tự động**, hạn chế tối đa giao dịch tay:

- **Không biên tập lại lời người nói.** Văn bản trong báo cáo trích từ lớp text máy đọc của file gốc; các câu trích dẫn trực tiếp (ví dụ *"An Agentic SOC is a system, not just a model"*) giữ nguyên nguyên văn, kể cả khi cách diễn đạt của diễn giả không chuẩn mực ngữ pháp.
- **Truy vết hai chiều tự động.** Mỗi khẳng định định lượng trong báo cáo đều tồn tại chuỗi truy vết: câu trong báo cáo → file `.txt` tương ứng → số trang/slide cụ thể trong file gốc. Bất kỳ ai cũng có thể kiểm chứng lại bằng công cụ tìm kiếm thông thường trên thư mục `extracted/`.
- **Phân tách dữ liệu và diễn giải.** Nội dung tường thuật (bài nói gì, vấn đề gì, kết quả gì) được tách khỏi phần đánh giá; báo cáo **không chấm điểm, không xếp hạng** chất lượng các bài trình bày, không bổ sung nhận định ngoài tài liệu nguồn.

- **Xử lý thống nhất bằng kịch bản.** Các phép biến đổi văn bản (chuẩn hóa dấu gạch ngang, tách danh sách, autolink URL, sinh mục lục và anchor) được thực hiện bằng kịch bản lặp lại được, áp dụng **cùng một luật cho mọi bài**, loại trừ thiên lệch do chọn lọc thủ công. Các kịch bản nằm trong quá trình làm việc và quy ước được ghi lại trong `AGENTS.md`.
- **Giới hạn dữ liệu nguồn được công bố minh bạch.** Tham luận #11 (Apple Secure Kernel) và #13 (MustangPanda/PlugX) không có slide công khai; báo cáo ghi nhận trạng thái này ở phụ lục thay vì phỏng đoán nội dung từ tiêu đề.
- **Đa số số liệu mang tính dẫn lại từ slide, không kiểm chứng độc lập.** Báo cáo ghi rõ nguồn mà mỗi slide tự trích (nếu có) để người đọc tự đánh giá độ tin cậy; đây là ranh giới trách nhiệm của một tài liệu tổng hợp, khác với một cuộc kiểm toán độc lập.

Phạm vi và giới hạn trách nhiệm

Tài liệu này là **bản tổng hợp thông tin** phục vụ tra cứu nội dung hội thảo, không phải văn bản pháp lý hay đánh giá chuyên môn. Mọi số liệu, tuyên bố kỹ thuật và kết quả nghiên cứu thuộc về tác giả của từng bài trình bày và nguồn mà họ trích dẫn.

Báo cáo được tổng hợp từ 25 bộ slide gốc (trong `Presentations/`) với phần trích xuất text tại `extracted/` và hình tổng quan tại `assets/figures/`. Xem `README.md` để biết cấu trúc dữ liệu và `AGENTS.md` để biết quy ước xử lý.