From Human-Led Cybercrime to Agent-Orchestrated Crime: A New Forensic Imperative

Editorial

From Human-Led Cybercrime to Agent-Orchestrated Crime: A New Forensic Imperative

Kaukab Jamal Zuberi

That model is beginning to change.

Artificial intelligence is moving beyond content generation towards agentic systems capable of planning tasks, using software tools, accessing external resources, making limited decisions and executing sequences of actions with reduced human supervision. When several specialised agents are connected through an orchestration framework, they can divide a complex objective into separate tasks, communicate results and modify their actions according to changing circumstances.

Consequently, electronic crime may no longer be conducted through a single program, account or operator. It may be executed through a coordinated framework involving a planning agent, reconnaissance agents, coding agents, social-engineering agents, data-analysis agents and agents responsible for infrastructure, communication or concealment.

Europol has already warned that artificial intelligence and automation are increasing the speed, scale and effectiveness of online fraud, social engineering and other forms of serious and organised crime. Its 2026 Internet Organised Crime Threat Assessment reports that generative AI is being used to tailor social-engineering activity and to accelerate and conceal fraudulent operations. Europol has similarly described AI as a catalyst that enables criminal networks to automate activities, target more victims and expand operations.

We must, however, avoid exaggeration. Available public evidence does not yet establish that autonomous multi-agent criminal systems are routinely conducting complete cybercrime campaigns without human involvement. What can be confirmed is that the technical capability is developing rapidly. In controlled research, teams of large-language-model agents have been used to investigate and exploit benchmarked real-world vulnerabilities, including previously unknown vulnerabilities. The research demonstrates how a planning agent can assign work to specialised sub-agents, overcome limitations encountered by individual agents and maintain progress across a complex task.

This capability has an important implication: the architecture used for legitimate automation, cybersecurity testing and business operations can also be adapted or misused for criminal purposes.

The Unit of Investigation Is Changing

Traditional digital forensics usually begins with a person, device, account, application, server or network connection. Investigators acquire the relevant media, recover artefacts, establish a timeline and attribute actions to users or processes.

Agentic crime requires an expanded investigative model.

The relevant unit of analysis may no longer be one device or one user. It may be an entire agentic workflow distributed across cloud services, application programming interfaces, local computers, containers, databases, communication channels and third-party tools.

A single criminal objective could generate evidence across several layers:

- the instructions given by the human controller;

- system prompts defining the behaviour of individual agents;
- the identity, role and version of each agent;
- messages exchanged between agents;
- plans generated by an orchestration component;
- tools and permissions available to each agent;
- application programming interface requests;
- tool-call arguments and returned results;
- retrieved documents and external data sources;
- temporary and persistent agent memory;
- code created or executed by agents;
- cloud, identity, network and endpoint logs;
- human approvals, interventions or overrides; and
- the final actions performed against victims or systems.

This evidence may be volatile, distributed and controlled by several service providers. Some records may exist only briefly unless logging has been enabled in advance. Other records may contain confidential information, personal data, credentials or privileged material. Investigators will therefore need the technical capability and legal authority to identify, preserve and correlate evidence across the complete agent chain.

Existing observability standards illustrate the type of records that can become relevant. OpenTelemetry's generative-AI conventions provide fields for agent identifiers, agent versions, conversation identifiers, system instructions, model requests, model responses, retrieval operations, tool definitions, tool-call arguments and tool results. These conventions were created for system monitoring, but the same categories may become highly valuable during forensic reconstruction.

Attribution Will Become More Complex

In conventional cases, investigators often ask: "Who performed this action?"

In agentic cases, that question must be divided into several parts:

Who established the criminal objective? Who configured the agents? Which agent formulated the operational plan? Which agent selected the target? Which agent called the relevant tool? Was the action authorised by a human, generated autonomously or triggered by another agent? Was an agent operating as designed, manipulated by an external instruction or compromised by a hostile agent?

These distinctions will matter when determining intention, knowledge, participation, causation and responsibility.

An agent may also produce actions that were not expressly anticipated by its operator. Conversely, a suspect may claim that an autonomous system acted independently. Investigators must therefore distinguish between genuine autonomous behaviour, negligent configuration, deliberate delegation and an attempt to use AI autonomy as a defence against attribution.

The evidential record must connect the original objective to the planning process, agent communications, tool execution and resulting harm. A final output alone will rarely be sufficient.

New Forms of Evidence Manipulation

Multi-agent systems create additional opportunities for evidence manipulation. A malicious or compromised agent may provide false information to other agents, alter an operational plan, inject instructions into shared memory, impersonate a trusted agent or cause another agent to execute an unauthorised tool call.

OWASP's guidance on agentic applications identifies risks associated with excessive agent authority, unverified inter-agent communication, goal manipulation, cascading failures and unsafe tool use. These risks demonstrate that investigators must consider not only whether an AI agent participated in an incident, but also whether the agent itself was deceived, compromised or used as an intermediary.

MITRE has also expanded its ATLAS knowledge base to address agentic AI and large-language-model threats. MITRE notes that systems capable of independent decision-making and interaction across environments create new opportunities for adversaries and introduce additional risks to organisational threat landscapes.

Forensic examination must therefore include the integrity of the agent framework itself. Investigators may need to determine whether prompts were altered, memory was poisoned, retrieved information was manipulated, credentials were misused or a legitimate agent was redirected towards an unlawful objective.

Investigative Agencies Must Upgrade Now

Investigative agencies cannot wait until autonomous or multi-agent crime becomes widespread before developing a response. Capacity building must begin while these systems are still emerging.

Digital-forensics laboratories should develop procedures for acquiring agent logs, orchestration records, cloud audit trails, application programming interface histories, vector databases, prompt repositories, tool configurations and containerised execution environments. Investigators should be trained to reconstruct agent-to-agent communications and correlate them with endpoint, network, identity and financial evidence.

Agencies will also require multidisciplinary teams. Traditional computer and mobile forensic examiners will remain essential, but investigations may additionally require expertise in artificial intelligence, cloud infrastructure, software development, identity management, model security, data science and financial investigation.

Legal frameworks and standard operating procedures must be reviewed to address the preservation and production of evidence held by AI-platform operators. Service providers should be encouraged—or, where appropriate, legally required—to maintain proportionate and privacy-conscious audit records that can identify agent activity without unnecessarily exposing sensitive user content.

Research has already begun examining artefacts produced by multi-agent platforms, including Microsoft's AutoGen framework. Such work confirms that agentic environments can produce identifiable forensic artefacts, but it also shows that this remains a developing field requiring further tool validation, repeatable acquisition methods and evidential standards.

A Call for Agentic-AI Forensics

The development of agentic systems does not make established forensic principles obsolete. Integrity, repeatability, documentation, validation, chain of custody and independent verification remain fundamental.

What must change is the scope of the investigation.

Electronic crime investigation must move from a primarily device-centred approach towards a workflow-centred and ecosystem-centred approach. Investigators must be able to reconstruct not merely what a computer executed, but how a collection of artificial agents interpreted an objective, divided responsibility, exchanged information, selected tools and produced an unlawful result.

The same technologies that strengthen defence, research and automation may also provide offenders with scalable and adaptive capabilities. The appropriate response is neither panic nor denial. It is preparation.

Investigative agencies, forensic laboratories, universities, prosecutors, regulators and technology providers must collaborate now to establish terminology, evidence-preservation requirements, forensic methodologies, training programmes and standards for agentic-AI investigations.

The next generation of electronic crime may not be committed by one offender operating one computer. It may be orchestrated by a human directing a network of specialised artificial agents. Our investigative capabilities must evolve accordingly.