Carnegie Mellon University
Software Engineering Institute

OBSERVATIONAL HUMAN–AI (OHAI): A DEFENDER ATTRIBUTION FRAMEWORK FOR DISTINGUISHING HUMAN VS. AI THREATS

Dustin D. Updyke
David Rossell
Shelly Fitzgerald

April 2025

[DISTRIBUTION STATEMENT A] Approved for public release and unlimited distribution.

Abstract

Artificial Intelligence (AI) is collapsing the cost, time, and skill barriers that once separated casual intruders from advanced persistent threats. For cyber defenders, the resulting evidence stream—rapid and adaptive—blurs the line between human and machine-enabled operations, undermining classic attribution methods and incident-response playbooks. This whitepaper introduces the *Observational Human–AI (OHAI) Attribution Framework*, a five-stage cycle of *Triage, Classify, Analyze, Profile,* and *Report* that enables at-rest and in-flight inspection to assign probabilistic confidence to the spectrum of attacker archetypes between human and fully autonomous AI attacks. We catalog observable AI indicators and demonstrate practical application through two publicly documented AI-enabled incidents. OHAI supplies defenders with definitions, analytic heuristics, and automation-ready data fields, enabling the faster discrimination of AI-driven threats, sharper predictive analytics, and more resilient human–machine defensive teaming. By operationalizing attribution, the framework aims to reduce misallocation of response resources, improve early-warning fidelity, and inform future toolchains that will operate against autonomous adversaries.

Introduction

Can humans clearly distinguish between human-executed and autonomous artificial intelligence (AI) cyber attacks? This paper explores how blue team defenders grapple with this question when analyzing evidence. We hypothesize that even experienced cybersecurity professionals face substantial uncertainty when attributing these activities to purely human or AI-driven adversaries. Defenders often operate without full visibility into attacker methods or intent and must make critical decisions under conditions of significant uncertainty while relying on incomplete data and inference.

[DISTRIBUTION STATEMENT A] Approved for public release and unlimited distribution.

Why does this human–AI distinction matter? Understanding whether an attack is human or AI driven may affect incident playbooks, response strategies, and strategic resource allocation. For defenders operating under uncertainty, misattribution can lead to ineffective defenses, wasted resources, or worse. It also brings broader implications for future workforce readiness: As AI evolves from mere tool to "teammate," how will defenders adapt to an environment where they often cannot see inside the attacker's methods?

We propose The Observational Human–AI (OHAI)[1] Attribution Framework for thinking about this distinction problem from the defender perspective. It is similar in spirit to many existing models, while incorporating the signals and methods that are well suited to AI detection. We also detail downstream considerations for incident response (IR), predictive analytics, and human–machine teaming assessment. Our goal is to help defensive security professionals adapt their reasoning, processes, and toolsets to successfully identify and mitigate evolving AI-powered threats. The call for research and training in this area is increasingly timely and relevant as incidents involving generative AI and sophisticated autonomous techniques rise in frequency and impact.

1. Problem Framing

Modern threats increasingly rely upon rapid automation and adaptive decision making. AI potentially expands these capabilities to an entirely new paradigm.

> *The threat posed by AI-driven cyberattacks significantly differs from traditional cyberattacks in terms of sophistication and scalability. AI allows attackers to target a wide variety of systems, each with unique vulnerabilities, all at once. What once required specialized expertise can now be done by novices using AI tools, making it easier for more people to launch attacks. This has led to a broader threat landscape where cyberattacks are not only more frequent but also potentially more dangerous due to their automated nature [Terrill 2024].*

This shift also increasingly complicates the traditional defender perspective, since the observable evidence of such threats that defenders are evaluating is often incomplete, ephemeral, and ambiguous. While existing intrusion frameworks outline the phases of a cyber attack, as well as the tactics, techniques, and procedures (TTPs) used within each, they do not provide significant clarity on how to specifically differentiate AI-orchestrated actions from human-driven ones. Additionally, many of the popular definitions of key defense terms and processes assume that defenders know many details about an attack, as if they are "on the inside," when from a purely defensive perspective, blue teams are making many decisions under uncertainty.

While piecing together evidence of any cyber attack is far from straightforward, attribution from that evidence has always been a murky and delicate task. With defenders lacking an "insider" view, they

[1] *OHAI* is a play on the Japanese greeting "ohayou"—meaning "good morning" or "hi," repurposed here as "oh, hi" to reflect recognizing AI-driven operations in the field.

[DISTRIBUTION STATEMENT A] Approved for public release and unlimited distribution.

can only evaluate fragments of logs, forensics, and indicators "from the outside looking in." When combining this difficulty with the additional attribution consideration of AI, some key points arise.

- The speed and adaptability of AI-driven attacks will potentially increase in a dramatic fashion. Machine learning (ML) algorithms allow attackers to automate tasks on a massive scale, making it difficult for defenders to confidently distinguish observed activity as led by AI or a human operator.
- There will be new AI-driven artifacts to evaluate and identify. With generative AI, it is trivial to produce highly polished phishing emails or polymorphic malware. Classic indicators of compromise may no longer suffice for these new types of artifacts.
- Defenders will continue to operate with only partial evidence. TTP-based attribution relies on linking technique usage to known threat actors. AI systems can combine or mutate known TTPs unpredictably. Many defenders remain unprepared to recognize or classify these mutations.
- As with traditional cyberattacks, AI will incorporate evasion tactics to operate clandestinely. Attackers can simulate human-like mistakes or intentionally behave "machine like" to mislead investigators. Attributing to AI vs. human will not just be a technical challenge but an epistemological one as well.

Traditional IR playbooks assume a relatively steady, human-paced, and at least somewhat predictable adversary. AI, by contrast, can adjust and reconfigure TTPs in near real time. It will be harder to predict new patterns of activity in such a rapidly shifting paradigm. As defenders attempt to piece together observable evidence, AI-enabled attacks will potentially be moving so quickly as to introduce new ambiguities into the challenge. Our approach is to reduce uncertainty by proposing a multi-layered, evidence-based framework attuned to AI indicators.

To summarize, knowing whether a cyber attack is human based or AI based is important for several reasons:

- **Response strategy and attribution**: Identifying the source of an attack (human or AI) can indeed inform response strategies. For example, AI-powered attacks may require automated countermeasures due to their speed and adaptability, while human-driven attacks may require more intelligence-driven attribution and engagement strategies. Of course, attribution to AI doesn't always mean the absence of a human component—AI attacks may still be initiated or guided by humans.
- **Predictive analytics**: AI attacks often rely on adaptive strategies that evolve based on their success, which makes them a valuable subject for improving predictive analytics. Analyzing how these attacks work can help to forecast future attacks. This applies not only to AI-driven attacks but also to blended threats where humans use AI tools, so it is advantageous to emphasize AI's role in enabling new attack paradigms.
- **Communication and collaboration**: Knowing whether an attack is human based or AI based can facilitate more effective communication and collaboration between IR teams, law enforcement, and other stakeholders. The distinction between human and AI attackers can influence legal responses, such as the application of cyber laws or treaties.

[DISTRIBUTION STATEMENT A] Approved for public release and unlimited distribution.

- **Security posture improvements**: By understanding the characteristics of AI-based attacks, organizations can refine their security posture to better defend against these types of threats, ultimately reducing the risk of future attacks. Organizations may prioritize AI-driven attacks differently due to their potential scale, speed, and automation capabilities.
- **Advanced cyberattack warning and attack assessment capabilities**: By improving the ability to detect whether an attack is AI driven or human driven, organizations can (1) adjust to the best counterstrategy in the short term and (2) improve their early-warning systems and harden their overall threat assessment processes in the long term. Discriminating between attacker profiles sharpens best triage procedures and helps prioritize response efforts. This distinction also enables more effective pattern recognition and predictive modeling, allowing defenders to foresee and counter advanced AI-enabled tactics or rapidly identify the hallmarks of human-led campaigns. Ultimately, these enhanced warning and assessment capabilities minimize the window of opportunity for attackers, reduce risk, and improve response readiness.
- **Assessment of human–machine teams (HMTs)**: As machines move beyond the role of a simple tool and are now sophisticated teammates that aid in decision making and defensive cyber operations, there will surely be a push for comprehensive HMT assessments. This will take the form of trainings and exercises for humans and the form of test, evaluation, validation, and verification (TEVV) for AI systems. And so, distinguishing humans from machines in terms of activity will be paramount.

By recognizing these indicators and understanding the importance of distinguishing between human-based and AI-based attacks, organizations can improve their defenses and respond more effectively to emerging cyber threats.

2. Objectives of This Attribution Framework

OHAI addresses the problem space laid out in the previous section by proposing clear definitions, investigative methods, analytical processes, and strategic implications. We also want its structure to serve as a foundation for thinking about automated tools that will operationalize AI attribution—this automation is sure to follow in this space if progress on attribution is successful. This begs the question, if we were building a tool today to automate the evaluation of evidence for AI attribution, how would we proceed?

While we stress that this is preliminary work, we believe there is value in starting the discussion now for this domain. The distinction problem seems an inevitable outcome in the continued coevolution of attack and defense within the cyber domain, as,

> *A prevalent type of co-evolution is that which is seen in predator-prey evolution where both sides evolve in terms of speed, stealth, camouflage, sense of smell, sight, and hearing as necessary to survive…the cybersecurity technological co-evolution that takes place between cyber attacker and defender, a process which clearly follows the predator-prey model more closely. Understanding this form of co-evolution enables defenders to position themselves strategically to get ahead of cyber threats [Willard 2015].*

[DISTRIBUTION STATEMENT A] Approved for public release and unlimited distribution.

If we believe attackers will gain some advantage by using AI, then we should also believe there is advantage in countering that advantage. The Department of Defense is clearly thinking in the direction of AI-led cyber attacks and defense, as retired Gen. Paul Nakasone stated,

> *'We're starting to challenge this idea of humans in the middle of the loop, and I also offer to you as we think about artificial intelligence models, think about cyber weaponry,' he said. 'How far are we talking to this idea of being able to create an agent that's going to move through your network, that's going to change based upon topology in the network, being able to evade the defenses that are there, choosing targets of the future?' [Starks 2025]*

Defenders having a counter to these types of AI-led attacks is the primary motivator for the OHAI effort.[2] Specifically, some of the areas where we seek to add value include

- **Clarifying key terms**: Clearly defining and differentiating "attack" and "AI-generated artifacts," among other general or vague terminology, will be valuable for defender operations. Establishing precise definitions will support consistent identification and classification of malicious activities involving varying degrees of AI automation or autonomy. This shared understanding is vital for coherent threat intelligence and informed defensive operations.
- **Outlining investigation concerns**: With definitions in place, being able to systematically identify forensic indicators and signals suggestive of AI involvement is a key next step. Observables such as adaptive speed, sophisticated polymorphism, real-time responsiveness, and strategic adaptability are characteristics distinguishing AI-enhanced or AI-driven attacks from purely scripted or manual operations. Providing defenders with clear guidance on these signals enhances their ability to detect, interpret, and respond appropriately.
- **Proposing a general framework**: OHAI should enable the classification of incidents into categories such as human, human–AI hybrid, or fully autonomous AI-driven attacks. By combining technical forensic analysis with contextual intelligence (such as historical patterns, threat actor profiling, behavioral analytics, etc.), the framework provides defenders with a structured, reliable mechanism to hedge cyber threats accurately.
- **Facilitating the automation of attribution**: We want to design operationalization into automated defensive tools and platforms. Clearly defined attributes, measurable observables, and structured inference methods create a foundation for automation via ML and AI-based detection systems. This automation capability enables rapid, scalable attribution and response, significantly reducing the human analyst workload while increasing detection speed, accuracy, and consistency.
- **Discussing follow-on strategic implications**: Improved attribution accuracy should affect broader cybersecurity strategies, including IR, threat forecasting, and resource prioritization. The framework helps defenders distinguish between isolated attacks and broader AI-driven campaigns, directly influencing strategic decisions about cybersecurity posture, resilience planning, and the deployment of defensive measures. By improving attribution capabilities, defenders gain

[2] This paper focuses on analyzing attacks from the perspective of the defensive team only.

[DISTRIBUTION STATEMENT A] Approved for public release and unlimited distribution.

clear insights into adversary intentions, strengths, and potential future threats, ultimately enhancing organizational cybersecurity readiness and response effectiveness.

In summary, this attribution framework aims to provide defenders with a structured, practical approach to identifying and responding to AI-driven threats. By clearly defining terms, systematically analyzing evidence, and enabling automation, organizations can significantly enhance their cybersecurity capabilities and resilience in an increasingly complex threat environment. Ultimately, through the adoption of this attribution framework, defenders can better prepare for and respond to the evolving challenges posed by AI-driven cyber threats.

3. Existing Literature and Case Studies

As recently as 2023, researchers had not confirmed that AI could autonomously conduct cyber attacks, but emerging evidence now indicates that this is no longer the case.

Mirsky organizes 32 capabilities of offensive AI and "offensive abuse of AI" into seven categories and then maps those to existing intrusion frameworks [Mirsky 2023]. They also interview cybersecurity experts and ask them to rate the offensive AI capabilities that represent the greatest risk to organizations. Respondents note their greatest concern revolves around social engineering attacks.

Research from Hamin highlights concerns about the use of generative AI in attacks, particularly in the phases of an attack focused on establishing initial access and persisting by using social engineering techniques [Hamin 2024]. They argue that large language models (LLMs) will provide less sophisticated attackers—those that lack the resources and expertise of advanced persistent threats (APTs)—with enhanced capabilities formerly available to only the most sophisticated attackers.

Fang provides detailed walk-throughs of how LLMs can be used to autonomously attack websites and exploit zero-day vulnerabilities [Fang 2024].

Other authors such as Happe, have argued that AI can be used to support penetration testing or help with the production of elements in a multi-stage attack [Happe 2023].

4. Key Terms and Perspectives on Observation, Evidence, and Evaluation

Terms such as "AI" and "autonomous" mean different things to different people in different contexts, so before we get started, we will define the core terminology used in relation to the OHAI framework:

- **AI**: We use the definition from the U.S. National Artificial Intelligence Initiative Act of 2020. While AI has advanced substantially in the intervening years, we feel that this definition is sufficiently generalized to retain its usefulness to defenders today.

  The term "artificial intelligence" means a machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations or decisions influencing real or virtual environments. Artificial intelligence systems use machine and human-based inputs to—(A) perceive real and virtual environments; (B) abstract such perceptions into models through

[DISTRIBUTION STATEMENT A] Approved for public release and unlimited distribution.

analysis in an automated manner; and (C) use model inference to formulate options for information or action [NAIIA 2020].

- **Cyber attack**: Our definition of cyber attack is drawn from the NIST glossary and some of its publications: "An attack, via cyberspace, for the purpose of disrupting, disabling, destroying, or maliciously controlling a computing environment/infrastructure; or destroying the integrity of the data or stealing controlled information" [NIST 2025]. We do not include mis/dis/mal-information attacks or social engineering attacks specifically within this paper, although we suspect that our approach generalizes to those scenarios.
- **Autonomous AI attack**: A cyber attack in which AI directs the details of the attack without the intervention of a human and without pre-programmed responses to predetermined stimuli or conditions.[3] This is distinguished from a scripting attack by the fact that even the most robust script bases its actions on a set of predefined inputs.
- **Human-Executed attack**: A cyber attack in which all elements are implemented by a human being without resorting to the use of AI or LLM tools. This includes attempting to log into a website as a legitimate user would using stolen login credentials.
- **AI-Mediated attack**: A cyber attack in which a human employs AI or LLM tools to assist in creating the attack. This might include having an LLM write a phishing message that the human sends, getting help from AI in writing a script that a human delivers and executes, or providing a rotation of command and control (C2) domains that a script automatically employs to obscure data exfiltration and C2 traffic. In these examples, the AI is used to support the human and is not able to make decisions regarding an attack.

Clarifying Observations: Artifacts at Rest and Activity in Motion

When defenders observe artifacts or activities, it is important to clarify precisely what they are examining. Accurate attribution depends heavily on distinguishing between different types of evidence and behaviors, which we broadly categorize as either artifacts "at rest" or activity "in motion."

- **Artifacts at rest**: These are static pieces of evidence, such as logs, files, malware binaries, scripts, or phishing messages. These can be stored in memory, on the file system, or aggregated within logging systems. Artifacts at rest exist independently of any immediate attacker actions and can generally be analyzed through static forensic methods. Examples include log analysis, code inspection, file system examination, and reviews of the linguistic or stylistic markers in phishing email content.
- **Activity in motion**: These involve dynamic, real-time behaviors indicative of active cyber attacks. Examples include machine-to-machine connections indicating potential lateral movement attempts, interactive C2 sessions, and changing behavior in response to defensive measures.

[3] Our experience is that fully autonomous AI attacks are difficult to track and report on in real-world events, but also to execute and study in controlled experimentation environments as well. This accounts for some of the subject's lack of existing detailed literature.

[DISTRIBUTION STATEMENT A] Approved for public release and unlimited distribution.

Activities in motion are typically observable only through real-time network analysis, behavioral analytics, or other defensive activity requiring the capture and analysis of real-time telemetry, network traffic flows, session interactions, and behavioral anomalies that differ markedly from "normal use" baselines.

Distinguishing between artifacts at rest and activity in motion provides a mechanism for further differentiating the evidence used for attribution. Defenders can investigate potential AI or ML-driven behaviors using more specific guidelines that we provide below.

Finally, and perhaps most importantly, no single observation or piece of evidence is likely to provide definitive attribution. To compound this, most adversaries will employ obfuscation and "counter-observability" tactics that include deliberate misdirection, randomization, and strategic employment of conventional attack scripts to conceal AI-driven elements. As a result, we emphasize the necessity of employing a multi-layered approach that integrates static and dynamic analysis to reliably attribute AI involvement and best understand its implications tactically and strategically.

5. Framework for Evidence Evaluation

Below is a straightforward, step-by-step framework that ties together (1) aforementioned potential AI indicators, (2) a general evidence evaluation process, and (3) key questions that arise during attribution. The OHAI framework is practical enough for day-to-day IR use while acknowledging the epistemological uncertainties inherent in cyber defense.

1. **Triage**: Gather and catalog all relevant artifacts and observations without bias. Most organizations are already doing this. We're not advocating anything specific for AI-led threats, and we're not asking for any additional logging and monitoring beyond what defending teams are already doing. Existing methods for quickly assessing and prioritizing incidents based on initial indicators, severity, and operational impact are an operational entry point for OHAI evaluation.
2. **Classify**: Here we simply categorize evidence by preliminary type—human-driven, hybrid human-AI, or fully autonomous AI-driven attack—using early forensic evidence and behavioral indicators. This classification guides subsequent analytical approaches.
3. **Analyze**: Conduct a deeper forensic and behavioral analysis of artifacts at rest (static binaries, code) and artifacts in motion (network traffic, live behaviors). Emphasize probabilistic reasoning to handle uncertainty and dynamically update confidence levels in AI attribution.

   The following is a non-exhaustive list of evaluation criteria for evidence indicative of AI influence:

General Indicators of AI

- **Speed or large-scale variations**: Continuous or huge changes in payloads or lures that exceed typical "script kiddie" randomization. A superhuman cadence of

[DISTRIBUTION STATEMENT A] Approved for public release and unlimited distribution.

behavior/action/response (or what we call the "super-OODA loop"[4]), as part of an automated decision-making process.

- **Adaptive or generative components**: Code or behavior that evolves in response to input, suggesting some feedback loop (e.g., dynamic reconfiguration based on the available target environment).
- **Advanced pattern recognition**: Unusual patterns of behavior or capabilities that systematically identify and exploit patterns (e.g., highly personalized phishing emails generated from user data, for which there is existing research).
- **Signs of ML libraries or artifacts**: References to TensorFlow, PyTorch, scikit-learn, or similarly other recognized ML libraries. May appear as imports, filenames, or suspicious binary blobs.
- **Environment sensing loops**: Logic that collects detailed target data and fine tunes attacks in near-real time, indicating a learning or optimization routine. Shows adaptive and evolving tactics to remain concealed from detection.
- **Exceptional efforts to remain clandestine in operations**: Notable effort to remain hidden or showing the sophisticated use of encryption and evasion techniques. Manipulation of data to cover tracks. Manipulation of environment to thwart human and defensive tool detection.
- **Elevated polish**: Lack of typical human error patterns.

Additional Static AI Indicators (for At-Rest Artifacts)

- **ML framework or model files**: Finding .pb (Tensorflow), .pt (Pytorch), .onnx (open-ml format) or other known ML-related files; imports referencing AI libraries; or hidden data that looks like neural network weights.
- **Unusual obfuscation (encoded parameters)**: Large blocks of encoded data or parameters—often net weights—versus typical hardcoded constants or simple scripts.
- **Data-driven logic modules**: Code or configurations that consume labeled or structured data (e.g., training sets, feature vectors), rather than rigid logic or known exploit libraries.
- **Evidence of AI-specific preprocessing**: Scripts or routines for data cleanup, dimensionality reduction, or feature engineering, hinting at an ML pipeline.

Additional Active AI Indicators (for In-Motion Activity)

- **Dynamic adaptability and payload mutation**: Malware or attack behaviors that pivot rapidly based on defender actions or changing network conditions—suggesting automated learning or reactive algorithms.

[4] An OODA loop is a decision-making cycle of *Observe, Orient, Decide, Act*, developed by military strategist John Boyd that emphasizes speed and adaptability in rapidly changing environments.

[DISTRIBUTION STATEMENT A] Approved for public release and unlimited distribution.

- **Sophisticated evasion**: Real-time detection and bypass of security controls (e.g., sandbox detection, anti-AV measures[5]) that show iterative self-improvement rather than fixed tactics.
- **Orchestrated or swarm-like coordination**: Multiple bots or agents coordinating in ways that indicate centralized AI-driven decision making (e.g., load balancing, target prioritization).
- **Feedback-driven personalization**: Observed use of target-specific data to adjust attacks in real time (e.g., refining phishing language or tailoring exploit attempts).
- **Fast and automated reconnaissance**: Large-scale scanning or data harvesting that appears systematic and adaptive, versus a single human's ad-hoc steps.

There may be other factors that we have not considered or those that will evolve as more research in this area is completed.

4. **Profile**: Once we have evaluations of evidence, we can begin to develop attacker profiles based on integrating technical, behavioral, and potential historical evidence. Like other frameworks, profiles inform defenders about attacker capabilities, motives, and TTPs, aiding predictive analytics and future defensive strategies.
5. **Report**: Document findings comprehensively, providing clear attribution outcomes, associated confidence levels, and strategic implications for stakeholders. Reporting supports informed decision making, enhances IR plans, and contributes to organizational cybersecurity maturity. This is analogous to intel operations aiding and assisting purely cyber operations.

By following this stepwise approach, defenders gain a structured lens for evaluating whether a threat is purely human, AI-assisted, or predominantly machine-driven.

In our research, we conclude that frameworks such as ATT&CK do not necessarily have a prescribed way of applying to activities or artifacts. Rather, the assumption is to apply the applicable TTPs as tags or classifications. For OHAI, we believe that providing a prescriptive way of thinking about this domain for defenders is essential.

[5] Anti-AV refers to techniques or tools designed to evade detection by antivirus software, often used by malware to remain clandestine during execution.

[DISTRIBUTION STATEMENT A] Approved for public release and unlimited distribution.

Figure 1: The OHAI Framework Cycle

We emphasize that OHAI is not a foolproof formula—attribution inherently involves many probabilistic judgments—but applying these systematic checks brings rigor and reduces guesswork as defenders begin to interrogate AI indicators.

6. Example Scenarios

To demonstrate the pragmatic value of OHAI, this section applies the framework directly to two recent and notable AI-enabled cyber incidents. By walking step-by-step through specific, documented threats, we illustrate how OHAI's systematic attribution methods enhance defenders' ability to classify and respond to sophisticated AI-enabled attacks. These cases highlight the framework's relevance in clarifying complex real-world scenarios and preparing defensive teams for the evolving AI-enabled threat landscape.

6.1 Deepfake Audio Fraud (2019)

In 2019, attackers leveraged AI-generated voice deepfakes to convincingly impersonate an executive, resulting in unauthorized financial transfers exceeding $240,000 [Damiani 2019]. This incident underscores the practical value of applying a structured attribution framework to AI-enabled cyber threats.

1. Triage: Defenders begin by isolating initial indicators—in this case, reports of suspicious voice calls, unusual financial transactions, and discrepancies in executive communications. Early identification of anomalous behavior triggers immediate mitigation steps, such as freezing related accounts or verifying transactions through alternative channels.

2. Classify: The incident might be classified as "at rest" regarding the audio recording and "in motion" for the real-time impersonation. Understanding that the attack involved voice deepfake

[DISTRIBUTION STATEMENT A] Approved for public release and unlimited distribution.

technology directs the investigation toward AI-generated social engineering rather than traditional malware or infrastructure compromise.

3. Analyze: Forensic analysis focuses on the features distinctive to the generated audio, perhaps including subtle irregularities that are detectable by specialized deepfake detection tools. Analysts should investigate the quality of the audio artifact, metadata, and behaviors inconsistent with typical human communication.

4. Profile: Using the insights gained, defenders can begin to build an attacker profile: the familiarity with organizational structure, the targeted selection of high-value individuals, and the exploitation of trust dynamics within the organization. This step integrates both technical forensics and contextual organizational analysis, improving attribution accuracy.

5. Report: A report captures key observations from above, provides attribution confidence levels, identifies vulnerabilities, and lists recommended improvements. This might include organizational policy changes, heightened awareness training against AI-enabled social engineering, and improved verification protocols for high-risk financial operations.

Applying the attribution framework to this incident clarifies defensive priorities and demonstrates the practical utility of systematic analysis in addressing emerging AI-driven threats.

6.2 DeepLocker (2018)

The 2018 DeepLocker proof-of-concept demonstrated malware that concealed payloads behind a neural network, triggering activation only under specific facial-recognition, geolocation, and voice-recognition conditions [Kirat 2018]. This sophisticated use of AI highlights the relevance of OHAI to advanced threats that leverage new methods of concealment.

1. Triage: Defenders might note suspicious executables that exhibit no immediately observable malicious behavior yet curiously evade traditional static or dynamic analysis. Initial triage involves recognizing that malware behavior is particularly conditional, prompting a path of forensic analysis rather than standard isolate-and-mitigate procedures.

2. Classify: The dormant malware binary artifact classification is “at rest” initially, transitioning to “in motion” upon activation. The conditional trigger is identified as the critical distinguishing factor, leading analysts to classify the incident as likely to be an advanced AI-enabled threat.

3. Analyze: Attempts to reverse-engineer the conditional logic triggers might employ an analysis environment that mimics the clues of activation conditions (e.g., facial recognition) to uncover the hidden payload. These detection approaches will likely need to shift from traditional signature-based methods to behavioral and contextual analyses.

4. Profile: Analysts begin to develop an attacker profile focused on technical proficiency, particularly in AI. This profile will also emphasize the attacker’s strategic goals: targeting, stealth, and evasion of known standard detection methods. This profile suggests a sophisticated threat actor, potentially a state-sponsored or highly specialized criminal group.

[DISTRIBUTION STATEMENT A] Approved for public release and unlimited distribution.

5. Report: The final report provides attribution insights, explicitly highlighting how AI-enabled stealth and conditional activation significantly complicated the detection and defense of this attack. Recommendations might include enhancing defenses with new behavioral analytics or adopting adversarial ML techniques. These may include revising standard IR playbooks to address threats specifically leveraging AI-enabled conditional payload activation.

These two examples underscore the growing sophistication and variety of AI-driven threats. Defenders are learning from these incidents to detect generative content and build intelligence on emerging threat actors and their TTPs.

7. Broader Implications for Cyber Defense

As AI-assisted threats evolve, organizations must reconsider their defensive postures. Successfully managing these new challenges requires a potential shift in IR strategies, predictive analytical approaches, and long-term planning. The following sections detail adjustments that defenders might consider in order to best counteract evolving cyber adversaries utilizing AI.

7.1 Incident Response

AI-assisted attacks demand accelerated identification and mitigation strategies due to their ability to rapidly adapt and modify tactics in real time. Traditional IR playbooks, designed around slower, human-paced attack timelines, may become insufficient. To combat this, teams may need to adopt new response strategies to effectively deal with AI-enabled attacks. In addition, continuous updating of evidence indicative of AI-generated attack scenarios enables ongoing readiness against evolving threats.

7.2 Predictive Analytics and Threat Hunting

Security products and IR teams are increasingly turning to ML to spot subtle network anomalies or to identify malicious code that rapidly changes signatures. The synergy of human analysts plus AI-based detection is key: advanced heuristics can flag suspicious content, while human expertise confirms or refines the suspicion.

7.3 Human–Machine Team Assessments

Organizations must assess how effectively AI tools integrate with human defenders. Human analysts bring context and intuition, while AI brings speed and pattern recognition. Teams that coordinate both effectively will fare better against adaptive AI adversaries. This includes establishing trust and explainability in AI-driven defense tools.

7.4 Strategic Planning

Organizations should adopt a future-proof perspective on AI threats. This includes

[DISTRIBUTION STATEMENT A] Approved for public release and unlimited distribution.

- **Training and workforce**: Upskilling teams to interpret ML-based forensic outputs and handle AI-driven incidents.
- **Legal and policy implications**: Clarifying liability and regulations for AI usage in offensive and defensive operations.
- **Resilience**: Working from the assumption that some attacks will succeed at machine speed, building robust containment and recovery mechanisms is imperative.

Addressing an evolving landscape of cyber threats requires organizations to reconsider their approaches to IR, analytics, teamwork, and strategic foresight to include this new human–AI distinction. By enabling processes, tools, and workforce strategies to consider the spectrum of AI involvement in a cyber attack, defenders enhance their ability to detect, mitigate, and respond to increasingly sophisticated cyber adversaries.

8. Conclusion

AI adds an intricate new layer to the defender's uncertainty. From phishing content generation to autonomous exploitation, it accelerates attacks while obscuring their origins. Yet defenders can still mitigate these threats by extending proven attribution models with AI-targeted analysis—and yes, many researchers still believe humans are the crucial piece in an effective defensive team. Sarker says, "…in terms of context understanding, intuition, accountability, and creativity in designing unique security solutions, the human element remains crucial" [Sarker 2024].

The OHAI framework proposed here offers a structured path for gathering, contextualizing, and synthesizing indicators of AI involvement. Moving forward, we suggest the following actions:

- **Practical implementation**: Security teams should adapt IR playbooks to account for AI-driven attacks, and train personnel on advanced forensics and generative content detection.
- **Ongoing research**: Academic and industry collaboration is needed to refine AI detection techniques, especially amid the arms race of generative AI.
- **Human–AI collaboration**: Defenders must embrace human–machine teaming, where each complements the other's strengths in speed, accuracy, creativity, and context interpretation.

As AI matures, so will AI-augmented intrusions. The community's response must be an equally agile adaptation, blending technical innovation with robust epistemic frameworks for attribution. Only then can defenders stay ahead of these rapidly evolving and often opaque threats.

[DISTRIBUTION STATEMENT A] Approved for public release and unlimited distribution.

Acknowledgements

The authors would like to thank the following for their advice and insight: Andrew Mellinger (CMU SEI), Dr. David Danks (UCSD), Tyler Brooks (CMU SEI), Dr. William Scherlis (CMU SEI), Dr. Shing-hon Lau (CMU SEI), John Yarger (CMU SEI), Chris Inacio (CMU SEI), and Greg Touhill (CMU SEI).

References

[Damiani 2019]
Damiani, Jesse. A Voice Deepfake Was Used To Scam A CEO Out Of $243,000. *Forbes*. September 3, 2019. https://www.forbes.com/sites/jessedamiani/2019/09/03/a-voice-deepfake-was-used-to-scam-a-ceo-out-of-243000.

[Fang 2024]
Fang, Richard; Bindu, Rohan; Gupta, Akul; & Kang, Daniel. *LLM Agents can Autonomously Exploit One-day Vulnerabilities*. arXiv:2404.08144v2. Cornell University. 2024. https://arxiv.org/abs/2404.08144

[Hamin 2024]
Hamin, Maia & Scott, Stewart. *Hacking with AI: The Use of Generative AI in Malicious Cyber Activity*. Atlantic Council Cyber Statecraft Initiative. 2024. https://www.atlanticcouncil.org/in-depth-research-reports/report/hacking-with-ai/

[Happe 2023]
Happe, Andreas & Cito, Jürgen. Getting pwn'd by AI: Penetration Testing with Large Language Models. Pages 2082–2086. In *Proceedings of the 31st ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering*. November 2023. https://dl.acm.org/doi/10.1145/3611643.3613083

[Kirat 2018]
Kirat, Dhilung; Jang, Jiyong; & Stoecklin, Marc. DeepLocker - Concealing Targeted Attacks with AI Locksmithing. Black Hat USA 2018. *IBM Website*. August 4, 2018. https://research.ibm.com/publications/deeplocker-concealing-targeted-attacks-with-ai-locksmithing

[Mirsky 2023]
Mirsky, Yisroel; Demontis, Ambra; Kotak, Jaidip; Shankar, Ram; Gelei, Deng; Yang, Liu; Zhang, Xiangyu; Pintor, Maura; Lee, Wenke; Elovici, Yuval; & Biggio, Battista. The Threat of Offensive AI

[DISTRIBUTION STATEMENT A] Approved for public release and unlimited distribution.

to Organizations. *Computers & Security*. Volume 124. January 1, 2023. https://doi.org/10.1016/j.cose.2022.103006.

[NAIIA 2020]
National Artificial Intelligence Initiative Act of 2020. H.R.6216. 116th Congress. 2020. https://www.congress.gov/bill/116th-congress/house-bill/6216

[NIST 2025]
NIST. Cyber Attack – Glossary. *NIST Computer Security Resource Center*. 2025 [accessed]. https://csrc.nist.gov/glossary/term/cyber_attack

[Sarker 2024]
Sarker, Iqbal H.; Janicke, Helge; Mohammad, Nazeeruddin; Watters, Paul; & Nepal, Surya. AI Potentiality and Awareness: A Position Paper from the Perspective of Human–AI Teaming in Cybersecurity. Pages 140–149. In *Intelligent Computing and Optimization.* December 2024. https://doi.org/10.1007/978-3-031-50887-5_14.

[Starks 2025]
Starks, Tim. Former NSA, Cyber Command Chief Paul Nakasone Says U.S. Falling Behind Its Enemies in Cyberspace. *CyberScoop*. February 22, 2025. https://cyberscoop.com/former-nsa-cyber-command-chief-paul-nakasone-enemies-cyberspace/

[Terrill 2024]
Terrill, Marshall. AI-Driven Cyberattacks More Sophisticated and Scalable, but ASU Expert Offers Solutions. *Arizona State University News*. October 18, 2024. https://news.asu.edu/20241018-science-and-technology-aidriven-cyberattacks-more-sophisticated-and-scalable-asu-expert

[Willard 2015]
Willard, Gerald N. Understanding the Co-Evolution of Cyber Defenses and Attacks to Achieve Enhanced Cybersecurity. *Journal of Information Warfare*. Volume 14. Issue 2. 2015. Pages 16–30. https://www.researchgate.net/publication/299394113_Understanding_the_Co-Evolution_of_Cyber_Defenses_and_Attacks_to_Achieve_Enhanced_Cybersecurity

[DISTRIBUTION STATEMENT A] Approved for public release and unlimited distribution.

Legal Markings

Copyright 2025 Carnegie Mellon University.

This material is based upon work funded and supported by the Department of Defense under Contract No. FA8702-15-D-0002 with Carnegie Mellon University for the operation of the Software Engineering Institute, a federally funded research and development center.

The view, opinions, and/or findings contained in this material are those of the author(s) and should not be construed as an official Government position, policy, or decision, unless designated by other documentation.

References herein to any specific entity, product, process, or service by trade name, trade mark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by Carnegie Mellon University or its Software Engineering Institute nor of Carnegie Mellon University - Software Engineering Institute by any such named or represented entity.

NO WARRANTY. THIS CARNEGIE MELLON UNIVERSITY AND SOFTWARE ENGINEERING INSTITUTE MATERIAL IS FURNISHED ON AN "AS-IS" BASIS. CARNEGIE MELLON UNIVERSITY MAKES NO WARRANTIES OF ANY KIND, EITHER EXPRESSED OR IMPLIED, AS TO ANY MATTER INCLUDING, BUT NOT LIMITED TO, WARRANTY OF FITNESS FOR PURPOSE OR MERCHANTABILITY, EXCLUSIVITY, OR RESULTS OBTAINED FROM USE OF THE MATERIAL. CARNEGIE MELLON UNIVERSITY DOES NOT MAKE ANY WARRANTY OF ANY KIND WITH RESPECT TO FREEDOM FROM PATENT, TRADEMARK, OR COPYRIGHT INFRINGEMENT.

[DISTRIBUTION STATEMENT A] This material has been approved for public release and unlimited distribution. Please see Copyright notice for non-US Government use and distribution.

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License. Requests for permission for non-licensed uses should be directed to the Software Engineering Institute at permission@sei.cmu.edu.

DM25-0527

Contact Us

Software Engineering Institute
4500 Fifth Avenue, Pittsburgh, PA 15213-2612

Phone: 412/268.5800 | 888.201.4479
Web: www.sei.cmu.edu
Email: info@sei.cmu.edu