

Cybersecurity in the Age of Artificial Intelligence: Emerging Threats and Defensive Strategies

Vaibhav Mishra¹, Abhinav Yadav², Adnan Siddiqui³, Devesh Katiyar⁴, Gaurav Goel⁵

^{1 2 3} Students, ^{4 5} Assistant Professors Department of Computer Science

Dr. Shakuntala Misra National Rehabilitation University, Lucknow, Uttar Pradesh, India.

Abstract - Honestly, we came into this expecting to write a fairly standard paper about AI improving cybersecurity. What ended up in front of us was a lot more complicated than that. The same ML tools defenders use to identify and counter attacks are, structurally, the exact same tools attackers use to build attacks that slip past detection. We trace that this uncomfortable reality start from Sommer and Paxson's 2010 critique [1] three problems they identified, that bafflingly are still unresolved through the threats that are active right now: adversarial examples defeating production scanners, data poisoning baked invisibly into models before they ever deploy, Emotet running for seven years without ever being technically beaten, voice cloning that took €220,000 from a CEO who did everything right, and LLMs that have essentially destroyed the economic barrier that kept targeted phishing in check. We look at four cases that stuck with us: UK Energy CEO fraud, the Microsoft Tay bot disaster, Emotet, and Pegasus, and then we try to make an honest argument about what the path forward looks like which, we would argue, is less about better algorithms and more about better governance, smarter human-AI collaboration, and institutions that have finally caught up to the technology they have deployed.

KEYWORD: Cybersecurity, Artificial Intelligence, Machine Learning, Adversarial Attacks, Deepfakes, Data Poisoning, Phishing, AI Governance.

1. INTRODUCTION

We will be upfront about something: this paper did not go where we thought it would. We started with the reasonable expectation that AI in cybersecurity meant defenders were finally getting the tools to pull ahead. What we found during our research, was tracing how things that are attack and defence, actually played out in real scenarios and it was messier and, in a way, more interesting than anticipated. The technology arrived and both sides picked it up at roughly the same time. That is the situation the field is actually in, and a lot of the commentary has not fully reckoned with it.

For most of computing history, there was at least a practical separation between offense and defense. Building a detection system took one kind of expertise; finding gaps in

it took another. The barrier to mounting a sophisticated attack was genuinely high. Organizations could invest in commercial tools and a reasonably staffed security team and feel like they were keeping pace. That was not a perfect situation, but the asymmetry mostly favoured defense, and that mattered.

Machine learning changed that dynamic in a very specific way. Take a phishing detection model as an example it reads emails, learns what malicious ones look like, and flags new ones accordingly. Now take a phishing generation model it learns what convincing emails look like and writes new ones that dodge those flags. Architecturally, those are the same system. You flip the training objective and the shield becomes the weapon. With both the offense and defense having same code and same model. The cost of building the attack version, once you have the defensive version, is close to nothing [12].

For a few years, defenders did get the better of it. Around 2015 to 2018 or so, something real was happening: security operations teams that were completely buried in alerts ten thousand a day, ninety percent noise finally had tooling that could sort through the pile intelligently. Investigations that took a week were finishing in a day. People were starting to feel like they were actually gaining ground. We think that period was real, not imagined. This AI technology was really working.

Then the same capabilities showed up on the attacker side, and things got complicated again. Phishing quality jumped dramatically, and the cost to produce it dropped. Malwares such as Emotet and Storm Worm started rewriting its own signatures between infections, which made signature-based detection look increasingly beside the point. Voice synthesis crossed a threshold where it could fool a careful person in a live phone call not in a lab but in an actual fraud. Both sides are now running on the same fuel, and there's no version of that where the advantage stays fixed on one side [12].

To be clear about what we are and are not arguing here: we are not saying AI has made cyber security a lost cause. It

has no, at least not yet. The argument is more specific than that. The framing that AI gives defenders a decisive edge was always too simple and the reason it has not lived up to that framing has less to do with the quality of the algorithms than with structural problems that algorithms cannot fix. The most important work the field needs to do right now is institutional more than it is technical, and that is an uncomfortable conclusion for a field that tends to reach for technical solutions first.

2. LITERATURE REVIEW

Reading the early ML-in-security literature is genuinely strange if you know how things turned out. There is this confident energy to these papers moving from benchmark results to sweeping conclusions without much friction, a shared assumption that techniques working so well in image recognition and spam filtering would port cleanly to intrusion detection. Looking back, the accuracy numbers look impressive right up until you ask what they were actually measuring.

Sommer and Paxson [1] published “Outside the Closed World” in 2010 and it remains the most honest piece of writing in the field. They were not arguing that ML was the wrong approach. They were arguing that three specific properties of the security domain break the assumptions that make supervised learning reliable elsewhere, and that the field had been quietly ignoring all three.

The first is class imbalance, and it is brutal once you see it. Real network attacks are a tiny fraction of real traffic sometimes one in a million events. A classifier that just calls everything normal gets 99.9% accuracy on any realistic test set, catches nothing, and looks great on paper. Standard evaluation metrics do not flag this because they were not designed with this kind of imbalance in mind, and researchers kept publishing results that looked like progress and would have been operationally useless.

The second is **concept drift** [11]. Most ML domains are stable cats keep looking like cats, handwritten digits do not change to fool scanners. Adversaries are not like that. They read the same threat intelligence that defenders publish, watch what gets detected, and update their techniques. A model trained on last quarter’s attacks is already partially blind to what is happening this quarter, and the model does not know it is blind it keeps outputting confident predictions either way.

The third one took us the longest to really sit with. And it was false positives and false negatives, in most ML applications, are roughly equivalent mistakes. In security

they are not even close. A false positive means an analyst spends half an hour chasing a non-event. A false negative means a breach goes undetected, maybe for days or weeks, even months, with consequences that can define a company’s year. Evaluation frameworks that treat these as equivalent are quietly optimizing for the wrong outcome.

Here is the thing that genuinely surprised us: we went looking for solution to those three problems in papers from 2015 but did not find them. Then checked 2019 papers, but same issues are still there. Read 2023 papers, and still no closure regarding these three issues. All three, essentially unaddressed, described in almost the same terms Sommer and Paxson used over a decade earlier. That is either a sign of unusually deep structural problems or a sign that the field has not asked hard enough questions about its own assumptions. We came to a conclusion based on our research that it is a cocktail of both.

The other piece of foundational work we kept returning to is **Goodfellow et al.** [2] from 2015. They showed that carefully crafted, imperceptibly small perturbations to an input can flip a neural network’s output entirely and with high confidence the famous panda-that-a-classifier-calls-a-gibbon (a kind of ape) result. The security version is direct: modify a malicious file at the byte level in ways that do not affect what it does at runtime, and watch it clear a trained scanner. **Anderson** [5] and **Demetrio et al.** [7] confirmed this works against actual production endpoint security products that real organizations are running and not just research prototypes.

3. WHAT WE’RE ACTUALLY DEALING WITH

A. Voice Cloning

Of all the threats we covered, voice cloning is the one we kept coming back to not because it is technically the most sophisticated edit, not really but because it attacks something so basic that nobody thought to build a defense against it. We trust familiar voices. We have done it our entire lives because there was never a reason not to. That lifetime of completely reasonable, automatic trust is now an attack surface.

Building a convincing deepfake video takes real resources: footage, compute, editing work. Building a voice clone that holds up in a live phone call takes maybe five minutes of target audio, and for any executive who has given a talk, done a podcast, or been on recorded earnings call that audio is on the internet right now available free of cost. The barrier to mounting this attack is essentially zero for anyone motivated to try [13].

Security awareness training has spent years teaching people to scrutinize emails because malicious emails carry detectable anomalies. A well-executed voice clone carries none [19]. There is nothing to scrutinize. The person receiving the call has no reason to question its credibility. Their identity-verification instinct fires, concludes this sounds like my boss, and they act on it. The instinct is not broken. It is just been made exploitable by a technology that arrived faster than any institutional defense for it did.

B. Adversarial Attacks on Scanners

The adversarial example finding does not stay politely in image classification. It transfers directly to malware detection. Small, targeted modifications to a malicious executable operating at the byte level, completely invisible to a human reviewer can push the file across a scanner's decision boundary from flagged to clean. The file still executes exactly what it was written to execute. It just looks, to the model, like it doesn't [2], [4], [7].

What makes this particularly frustrating from a defensive standpoint: there's often nothing to find on manual review. The scanner cleared it. An analyst looking at the file finds nothing unusual. The malicious behaviour only surfaces at runtime. By then, it's too late to have helped.

C. Data Poisoning

If voice cloning is the loudest attack on this list, data poisoning is the quietest and that quietness is most of what makes it dangerous. The attack does not look like an attack while it is happening. It happens upstream of everything, often months before deployment. An adversary who can influence the training pipeline, or who can affect how training labels get assigned, can build specific blind spots directly into the model from the start [16].

The finished model is fine by every metric the deployment team has. It passes validation. It performs correctly on test data. It handles the vast majority of production inputs without issue. It fails silently, predictably, and only on the exact scenarios the attacker designed it to fail on. By the time someone figures out what happened, the poisoned model may have been making real decisions for months. This is the attack that is hardest to catch because it does not announce itself.

D. Adaptive Malware The Emotet Story

We'd read about Emotet before this paper. Reading the actual timeline of it was still something. It ran from 2014 to 2021 about seven years. In that time, it got caught,

contained, and declared gone repeatedly. Security teams would write up how they'd stopped it, publish indicators of compromise, document the detection rules, and then it would come back a few weeks later in a form those rules did not cover.

Because Emotet was reading those writeups. It went dormant in sandbox environments to avoid automated analysis. It rewrote its signatures between infections. It monitored published threat intelligence from security vendors and updated its evasion logic in direct response treating the security industry's own research as a real-time improvement service [6]. Teams documenting their containment work were, without knowing it, contributing to its next iteration.

What finally stopped it was Europol, in January 2021, coordinating a physical seizure of its server infrastructure. Not a detection breakthrough. Not a better model. Law enforcement seizing hardware. The technical problem was never solved. It was bypassed.

E. LLMs and the Phishing Assumption That Broke

Every phishing training programme ever designed rests on one assumption: that personalised, contextually accurate, convincingly written phishing emails cost something to produce. That cost created the detectable artifacts people were trained to spot slightly off phrasing, generic greetings, implausible urgency. The expense was the constraint, and the artifacts were the tells.

Large language models erased that constraint. The cost of generating a phishing email that references a real project the target is working on, uses accurate internal terminology, and is written in fluent professional prose with none of the traditional red flags, that cost is now essentially zero [14]. We ran this ourselves during the research. Gave a widely available model a fake company, a job title, a plausible scenario. What came back would have given us pause if we had not written the prompt. The assumption the entire training infrastructure was built on is gone. Most programmes have not caught up.

IV. THREE CASES THAT STAYED WITH US

A. UK CEO Voice Fraud, 2019

In August 2019, the CEO of a UK energy company received a phone call from someone who sounded exactly like his boss at the German parent company. Same accent. Same rhythm. Same choices of words. Same particular register of urgency

that person used when something needed to happen quickly. He was told to transfer €220,000 to a Hungarian supplier that day, part of an acquisition. He assessed the call carefully. The voice sounded right. The story was plausible. He made the transfer.

The money moved through Hungary to Mexico within hours and was gone. The voice was synthetic [19].

We want to be precise about what happened here. The CEO did not cut corners. He performed exactly the verification any reasonably cautious person would perform he assessed the available identity signal, which was the voice, and it checked out. The attack worked not because he was careless but because the cognitive shortcut he used- this sounds like someone I know, so it probably was being exploited by a technology that arrived before any institutional defense for it did. That shortcut was perfectly reliable for all of human history until recently. It isn't anymore.

B. Microsoft Tay, 2016

Microsoft launched Tay in March 2016 as a chatbot that would learn and grow through conversations with Twitter users. The idea was a system that developed a genuine personality through real interaction. It lasted sixteen hours.

A coordinated group of users identified the feedback mechanism, how Tay updated its outputs based on what people said and they spent the afternoon methodically cycling harmful content through it until the model reproduced it confidently. Tay did exactly what it was built to do. The design had not considered adversarial users operating in coordination, which is to say it had not considered the internet as it actually is.

People usually take a content moderation lesson from Tay. We think the deeper point is that what happened was real-time data poisoning in production. No access to training infrastructure required. No technical sophistication beyond identifying the feedback loop. Just coordination. Any learning system that updates on user interaction faces some version of this. This led to AI organizations introduce filters and safeguards to protect the integrity of their model.

C. Pegasus

NSO Group's Pegasus is documented in careful detail by Citizen Lab at the University of Toronto, and the documentation is sobering [8]. The attack exploited a vulnerability in how iMessage parsed a specific file type. The result: complete device compromises the moment a message

arrived before the screen was looked at, before anything was tapped, with nothing visible to indicate anything had happened.

No link to avoid clicking. No attachment to decide against opening. No action available to the target that would have changed the outcome [9]. The device was compromised at message delivery. This is a fundamentally different category of threat from anything user awareness training addresses, because user awareness training assumes the user has a decision to make. At the zero-click level, they don't.

V. WHY FIFTEEN YEARS OF RESEARCH HAVEN'T FIXED THIS

The fair question at this point is: why not? Why do the exact problems Sommer and Paxson named in 2010 still show up in current papers? The researchers in this field are good. The answer is not a lack of effort. The answer is that the most important barriers are not technical, and you can't debug your way past a structural problem.

The data-privacy tension is one of these. Effective security AI runs on large, longitudinal datasets of real behavioural data. Privacy law, for entirely legitimate reasons, pushes in the opposite direction: collect the minimum data, retain it briefly and limit the use of that data. Both positions are defensible. Together, they're in direct conflict, and every security team is quietly making a judgment call somewhere in the middle, usually without documenting it, hoping it does not get tested in a way that makes it visible.

Embedded bias is the problem we suspect most organizations are not thinking about at all. ML models do not just learn signal they learn the statistical patterns in their training data, including the biased ones. If historical data reflects patterns where certain behaviours or profiles got flagged more often for whatever historical reasons those patterns exist and the model learns to reproduce them. The system can be accurate to its training distribution and systematically unfair in ways that are hard to see without specifically looking. We'd be surprised if more than a small fraction of organizations are auditing their deployed security models for this [17], [18].

The explainability gap bothers us most on an operational level. The models with the best detection performance are almost always the opaquest about why they produced a given output. An analyst who gets a high-confidence alert with no supporting reasoning has to decide whether to escalate or dismiss solely on the basis of a number. When the model is wrong, nothing in the output suggests it might be. This drives alert fatigue in a direct, rational way: analysts

learn to discount confident alerts because confident alerts are sometimes wrong and they cannot tell which ones. That's not an irrational response. It's the correct update given the information available [10], [15].

The economic asymmetry is probably the deepest problem and the furthest from any technical solution. Build one evasion technique, run it against thousands of organizations at near-zero marginal cost. Each of those organizations independently has to find it, understand it, and respond usually without knowing the others are dealing with the same thing. The total resource investment is on the defensive side and scales with the number of defenders. The attacker's cost stays flat. Better AI on the defensive side helps at the margins. It does not change that fundamental math [1], [12].

VI. WHAT'S ACTUALLY HELPING

We want to be honest that there are genuine advances here, because a paper that only catalogues problems isn't being fully accurate either. Some things are working.

SHAP (SHapley Additive exPlanations) and **LIME** (Local Interpretable Model-agnostic Explanations) are the tools which explain why a ML model took that specific decision when they take it and they do not reveal the entire process it takes to generate that response as it is lengthy and complex instead, they provide the list of inputs used and how the model's reasoning worked to reach this output. They have reached real practical utility in security operations. Not because they make models transparent in any deep sense, but because they provide **feature-level reasoning** for specific outputs. The shift from "94% confidence: threat" to "94% confidence: threat, driven primarily by features X and Y" is a small technical change and a meaningful operational one. It gives the analyst something to verify, something to investigate, a reason that can be written up and explained [10], [15].

The human-AI pairing that actually works is not the one that gets most of the marketing. The usual pitch is autonomous detection, AI that removes human judgment from the loop in the interest of speed. But what we see working is more like a sensible division of labour: AI handles what it is genuinely better at: continuous monitoring, consistent pattern matching, maintaining detection libraries, never getting tired at 3 am while humans handle what they're genuinely better at understanding what a pattern means in this specific environment, judging how serious it actually is, deciding what response is proportionate. Remove the AI and defenders get buried. Remove the humans and the

system makes confident errors with nobody to catch them [21].

The EU AI Act (2024) [20] matters not because regulation automatically produces good security outcomes but because it converts compliance from optional to mandatory, creating legal accountability for high risk AI. The voluntary best-practice frameworks that preceded it were useful for organizations already motivated to act. For the ones that were not, voluntary meant ignored. When real consequences attach to non-compliance, the calculus changes.

Post-quantum cryptography is the issue that worries us most relative to how much attention it is getting. NIST finalised its post-quantum standards in 2024. Most organizations have not started migration in any serious way. The threat is not immediate in the short term, probably but migration takes years even when it's prioritised, and data encrypted today can be stored and decrypted later when quantum hardware arrives. Being wrong about the timeline is not recoverable.

Autonomous response is the open problem we genuinely could not resolve. The speed argument is real but some attack categories move faster than human response time and automated action is necessary. But automated systems behave predictably, and predictable behaviour is exploitable. If an attacker can predict that a specific input triggers a specific automated response, they can trigger that response deliberately as an attack. We've spent time on this and do not have a satisfying answer. As far as we can tell, neither does the current literature.

The accountability question is the one we think will produce the most visible crisis the soonest. AI-driven security systems are already making decisions about real people- blocking access, flagging accounts, triggering incident response processes that affect careers. When something goes significantly wrong, the question of who is responsible is legally unclear in most jurisdictions. That ambiguity diffuses accountability in ways that make no one actually responsible for systematic errors. That gap will be tested.

VII. CONCLUSION

We set out to write about technology and ended up writing mostly about institutions. That probably reflects where the field actually is more accurately than we expected going in.

The algorithms are good. They have improved substantially over the fifteen years this paper covers. SHAP,

LIME, human-AI collaboration frameworks, and the movement toward enforceable governance represent real progress. We do not want to undersell that.

What has not kept pace is the infrastructure around the algorithms. Governance that matches the speed of deployment. Accountability frameworks that cover decisions these systems are already making about real people. Post-quantum readiness treated as a current priority rather than a future problem. The organizational capacity to audit deployed models for bias. Actual answers to questions about autonomous response that currently do not exist.

AI has made cyber security more powerful on both sides, more consequential when things go wrong, and faster-moving than any institutional process was designed to handle. It has not made defence easier. The organizations navigating AI-era threats best are not necessarily the ones with the most capable models they are the ones that pair technical capability with genuine governance, invest in human judgment alongside automated detection, and treat institutional readiness as a security requirement rather than a compliance checkbox.

Both sides will keep getting better tools. The question is whether governance, accountability, and institutional readiness improve at the same rate. Right now, honestly, they are not. That is the gap that actually matters.

REFERENCES

- [1] R. Sommer and V. Paxson, "Outside the closed world: On using machine learning for network intrusion detection," in Proc. 2010 IEEE Symposium on Security and Privacy, Oakland, CA, May 2010, pp. 305–316.
- [2] I. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," 2015, arXiv preprint arXiv:1412.6572. [Online]. Available: <https://arxiv.org/abs/1412.6572>
- [3] W. Hu and Y. Tan, "Generating adversarial malware examples for black-box attacks based on GAN," arXiv preprint arXiv:1702.05983, 2017.
- [4] N. Carlini and D. Wagner, "Audio adversarial examples: Targeted attacks on speech-to-text," in Proc. IEEE Security and Privacy Workshops (SPW), San Francisco, CA, 2018, pp. 1–7.
- [5] H. Anderson, S. Woodbridge, and B. Filar, "DeepDGA: Adversarially-tuned domain generation and detection," in Proc. ACM Workshop on Artificial Intelligence and Security (AISec'16), Vienna, Austria, 2016, pp. 13–21.
- [6] B. Biggio and F. Roli, "Wild patterns: Ten years after the rise of adversarial machine learning," Pattern Recognition, vol. 84, pp. 317–331, Dec. 2018.
- [7] L. Demetrio, S. Coull, B. Biggio, G. Lagorio, A. Armando, and F. Roli, "Adversarial EXamples: A survey and experimental evaluation of practical attacks on machine learning for Windows malware detection," ACM Transactions on Privacy and Security, vol. 24, no. 4, pp. 1–31, 2021.
- [8] B. Marczak, J. Scott-Railton, S. McKune, B. A. Razzak, and R. Deibert, "Hide and seek: Tracking NSO Group's Pegasus spyware to operations in 45 countries," The Citizen Lab, Munk School of Global Affairs, Univ. of Toronto, Toronto, Canada, Research Rep., Sep. 2018. [Online]. Available: <https://citizenlab.ca/2018/09/hide-and-peek-tracking-nso-groups-pegasus-spyware-to-operations-in-45-countries/>
- [9] C. Cimpanu, "Apple says NSO Group's zero-click iMessage exploit targeted journalists and activists," ZDNet, Sep. 13, 2021. [Online]. Available: <https://www.zdnet.com>
- [10] A. A. Chandio, N. Masood, A. Iqbal, and M. A. Tahir, "Explainable artificial intelligence (XAI) in cybersecurity: A comprehensive survey," IEEE Access, vol. 11, pp. 45836–45853, 2023.
- [11] M. Campos, J. J. Maestre Vidal, and A. F. Skarmeta, "Evaluating the impact of concept drift on machine learning-based network intrusion detection," IEEE Access, vol. 8, pp. 121567–121584, 2020.
- [12] N. Kaloudi and J. Li, "The AI-based cyber threat landscape: A survey," ACM Computing Surveys, vol. 53, no. 1, article 20, pp. 1–34, Feb. 2020.
- [13] Y. Mirsky and W. Lee, "The creation and detection of deepfakes: A survey," ACM Comput. Surv., vol. 54, no. 1, pp. 1–41, Jan. 2021.
- [14] J. Hazell, "Spear phishing with large language models," arXiv preprint arXiv:2305.01247, May 2023. [Online]. Available: <https://arxiv.org/abs/2305.01247>
- [15] S. M. Lundberg and S. Lee, "A unified approach to interpreting model predictions," in Advances in Neural Information Processing Systems (NeurIPS), vol. 30, Long Beach, CA, 2017, pp. 4765–4774.
- [16] M. Jagielski, A. Oprea, B. Biggio, C. Liu, C. Nita-Rotaru, and B. Li, "Manipulating machine learning: Poisoning attacks and countermeasures for regression learning," in

Proc. 39th IEEE Symposium on Security and Privacy, San Francisco, CA, 2018, pp. 19–35.

- [17] R. Binns, "Fairness in machine learning: Lessons from political philosophy," in Proc. Conference on Fairness, Accountability and Transparency (FAT*), New York, NY, USA, 2018, pp. 149–159.
- [18] N. Papernot and P. McDaniel, "Privacy and security in the age of machine learning," IEEE Security & Privacy, vol. 16, no. 3, pp. 56–59, May–Jun. 2018.
- [19] D. Statt, "Fraudsters used AI to mimic a CEO's voice in unusual cybercrime case," The Wall Street Journal, Aug. 30, 2019. [Online]. Available: <https://www.wsj.com/articles/fraudsters-used-ai-to-mimic-ceos-voice-in-unusual-cybercrime-case-11567157402>
- [20] European Parliament, "Regulation (EU) 2024/1689 of the European Parliament and of the Council on Artificial Intelligence (Artificial Intelligence Act)," Official Journal of the European Union, L series, Jun. 2024. [Online]. Available: <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32024R1689>
- [21] Gartner Inc., "AI for cybersecurity: Key use cases and recommendations for security leaders," Gartner Research, Stamford, CT, USA, White Paper, 2022.
- [22] Kaspersky Lab, "IT threat evolution in Q1 2023: Statistics," Kaspersky Securelist, Moscow, Russia, Quarterly Threat Report, Apr. 2023. [Online]. Available: <https://securelist.com/it-threat-evolution-q1-2023-statistics/109870/>