provadis Hochschule

Whitepaper – Prof. Dr. Volker Scheidemann, July 2025

The Good, the Bad and the Ugly: The Impact of AI on Information Security

1. Introduction: Artificial Intelligence Meets Cybersecurity

Artificial Intelligence (AI) has evolved from rule-based expert systems to powerful general-purpose tools capable of language generation, reasoning, code writing, and real-time anomaly detection. Especially since the release of OpenAI's GPT models, AI systems are increasingly being embedded across industries – including information security.

INHALT

From a technical perspective, current AI systems are largely powered by machine learning (ML) and deep learning (DL), with models trained on vast datasets of language, images, source code, or telemetry. Among these, Large Language Models (LLMs) such as GPT-4, Claude, Gemini, and LLaMA stand out for their ability to generate human-like text, reason over complex instructions, and write working software code – capabilities that defenders and attackers alike are now leveraging.

Several firms and organizations pictured of a "paradigm shift": AI would become both a force multiplier for cyber defense – and a weapon system for threat actors [1, 2]. A 2025 joint report from OpenAI, Microsoft, and several national agencies confirmed that state-sponsored attackers were experimenting with generative AI to write or debug malware, perform reconnaissance, and generate phishing content [3].

The double-edged nature of AI is now a permanent feature of cybersecurity. This paper explores both sides: the positive impact of AI on defending digital systems, and the growing risks posed by AI-assisted cybercrime – with a focus on real-world cases, the DACH region, and scientifically credible sources.

2. How AI Strengthens Information Security

2.1 Threat Detection and SOC Automation

One of the most transformative uses of AI in information security is in the Security Operations Center (SOC). AI agents – particularly ML-based systems – now analyze massive telemetry streams in real-time to detect anomalies, correlate alerts, and flag malicious behavior.

IBM's "Autonomous Threat Operations Machine" (ATOM) platform is an example of this trend: multiple generative AI agents are used to monitor incoming alerts, correlate logs, query threat intelligence databases, and even recommend mitigation strategies [4]. In practice, this enables response times measured in minutes rather than hours.

A 2024 study by IBM and Ponemon showed that organizations using AI and automation saved an average of USD 2.22 million per data breach, and reduced time to contain an incident by 108 days [5].

2.2 Behavior-Based Anomaly Detection

Traditional rule-based security systems often fail against sophisticated attacks or zero-day exploits. AI enables behavior-based detection: systems learn what "normal" traffic or user activity looks like and raise alarms when statistically significant deviations occur.

User and Entity Behavior Analytics (UEBA) systems, powered by ML, detect suspicious lateral movement, impossible travel, or privilege escalation by comparing user behavior across time and context [6]. These techniques now form the basis of many enterprise SIEM solutions and XDR platforms.

2.3 Threat Intelligence and Pattern Recognition

AI also strengthens threat intelligence. Generative and discriminative models alike are used to:

- Parse malware repositories for shared indicators of compromise
- Detect new phishing domains via DNS and WHOIS patterning
- Forecast attacker infrastructure by analyzing dark web chatter

For example, IBM's Predictive Threat Intelligence module draws from over 100 data sources, correlating malware strains and TTPs using natural language processing (NLP) and clustering techniques [4].

In 2024, Google's DeepMind also released SecPaLM, a fine-tuned LLM specifically trained for security log analysis and threat classification, capable of outperforming human analysts in phishing detection [7].

3. Generative AI in Active Defense and Prevention

3.1 AI for Vulnerability Discovery and Penetration Testing

AI is increasingly being used proactively: to scan for misconfigurations, simulate attacker behavior, and identify vulnerabilities before adversaries do. Tools like Microsoft's Security Copilot, IBM's ATOM Red Team Agent, and the open-source AutoRed system can now:

- Crawl infrastructure using NLP-augmented search queries
- Simulate privilege escalation paths
- Recommend mitigations in natural language

Morgan Stanley notes that "organizations are now targeting their own infrastructure using AI-driven red teaming to find weaknesses before attackers do" [8].

In a proof-of-concept published in 2024 by the MITRE Corporation, an LLM-assisted red team uncovered 37 % more exploitable paths in Active Directory environments than traditional automated tools [8].

3.2 Patch Management and Configuration Hardening

Another field being revolutionized is AI-assisted patching and remediation. Generative models trained on code repositories (e.g. Codex, StarCoder, CodeWhisperer) are now used to:

- Suggest configuration fixes for insecure settings
- Auto-generate patches for open-source libraries
- Analyze CVEs and summarize exploitability risks

For instance, SAP's proposed AI security framework integrates AI to scan ABAP code for insecure patterns and recommend remediations in development environments – reducing vulnerability exposure before deployment [9].

Furthermore, Palo Alto Networks' Cortex Xpanse platform uses AI to continuously scan exposed services and alert organizations to shadow IT and insecure configurations [9].

4. Case Studies: Defensive Success with AI

4.1 AI Prevents Data Exfiltration Attempt

In late 2024, a major European pharmaceutical company detected an anomalous download from a research subnet. An AI-enhanced UEBA system flagged the behavior due to:

- Login from an unusual IP address
- Data volume spike 300 % above baseline
- Off-hours activity from a research intern's account

Investigation revealed compromised VPN credentials. Thanks to the real-time alert, exfiltration was blocked, and incident response began within 12 minutes – saving an estimated €2.4 million in IP loss [10].

3.3 Natural Language Interfaces for Cybersecurity Operations

AI also lowers the barrier for security operators and junior analysts by translating complex technical actions into natural language prompts. For example:

"Show me all traffic anomalies on port 443 in the last 24 hours" → returns log segments with anomalies, risk scores, and GeoIP mapping.

This paradigm – sometimes referred to as conversational security ops – is now supported in platforms like Sentinel Copilot, IBM ATOM, and Google Chronicle. OpenAI's GPT-4 has been tested experimentally into SIEM platforms, demonstrating capabilities such as:

- Converting firewall logs into summaries
- Suggesting regex patterns for detection rules
- Explaining potential attack paths in plain English

These advances not only improve productivity but reduce human error, which remains a major factor in breaches.

4.2 SOC Copilot Resolves Alert Storm

During a large-scale phishing campaign in April 2025, a financial institution's SOC received 12,000+ alerts in one hour. Using IBM ATOM's AI triage module, 98.7 % of these were automatically correlated to a single phishing campaign, traced to three domains, and marked for blacklisting.

SOC response time dropped from an average of 90 minutes per case to under 15 minutes, with no escalation to manual analysts needed for 95 % of alerts [11].

5. Summary: The "Good" in AI for InfoSec

AI-driven cybersecurity is more than a buzzword – it's now a critical defense mechanism. From detection to triage, prediction to prevention, and even internal testing, AI may bring improvements:

- Faster detection and containment
- Fewer false positives
- Scalable threat intelligence
- More secure code before deployment
- Natural language interfaces that reduce SOC fatigue

These benefits are especially valuable in an era of staffing shortages in cybersecurity – with over 4 million unfilled positions globally as of 2025 [11].

6. The Bad or How AI Is Being Abused: Risks and Threats

6.1 AI-Powered Phishing and Social Engineering

The use of LLMs like GPT-3.5/4, Claude, and WormGPT for automated phishing has already moved from theory to practice. Unlike generic spam, AI-generated phishing messages can be:

- Context-aware (e.g., referencing recent transactions or known colleagues)
- Grammatically perfect (even in multiple languages)
- Highly scalable (generated in seconds for hundreds of victims)

A 2024 FBI alert warned of "AI-generated phishing emails impersonating internal communications, invoices, and even human resource notices" [12]. In many of these campaigns, the attacker used LLMs to automatically craft messages tailored to job roles or industries.

In one 2023 incident, Microsoft reported that the APT group Storm-0539 used LLMs to generate customized phishing emails for MFA bypass campaigns, successfully compromising cloud tenants [13].

Check Point Research similarly found that LLMs were being used to generate smishing (SMS phishing) messages with embedded malicious links or fake delivery notices [14].

6.2 Deepfakes and Voice Cloning

AI-generated deepfake audio and video has become a real operational threat.

- Video meetings with fake participants (using avatars) were used to obtain internal documents [12, 15]
- A case reported by the British newspaper The Guardian in 2023 involved a Hong Kong-based company that lost over £20 million after receiving deepfaked Zoom calls with realistic avatars of its executives [16].
- Morgan Stanley notes that "AI-generated voice and video content is now indistinguishable from real human output to the untrained observer" [1].

6.3 AI-Generated Malware and Exploit Automation

While most AI models have safeguards to block malicious queries, jailbroken or maliciously trained variants are now circulating on dark web markets. Examples include:

- WormGPT and FraudGPT – modified LLMs trained explicitly to assist in malware writing, password cracking, and social engineering [14]
- ChatGPT clones with "no ethical filters," offered as subscription services

In academic settings, researchers at the University of Sheffield and University of Surrey demonstrated that commercially available LLMs could generate polymorphic malware with trivial prompts [17].

In practice, AI is being used to:

- Obfuscate existing malware variants
- Generate new payloads for credential theft
- Write code for privilege escalation exploits
- Assist in reverse engineering or debugging

6.4 Adversarial Attacks on AI Systems

The irony of AI in InfoSec is that AI itself becomes a target. Attackers are increasingly exploiting vulnerabilities in ML models through:

- Prompt injection: Manipulating input to steer an LLM into producing harmful output
- Data poisoning: Injecting malicious examples into training data to bias outcomes
- Adversarial examples: Subtle changes in input (e.g., pixels in an image or log formatting) that fool classifiers

A well-publicized example from 2023 involved adversaries tricking an email classifier to mark phishing messages as "safe" by encoding text with invisible Unicode characters [18].

The Good, the Bad and the Ugly: The Impact of AI on Information Security

7. The Ugly: The Growing Abuse Economy

Several criminal marketplaces now offer AI-as-a-Service:

- Phishing-as-a-Service with dynamic LLM prompts
- Voice-cloning APIs for use in vishing campaigns
- Chatbot-based customer scams mimicking real brands

Several sources can be found documenting the use AI to produce child sexual abuse images [17, 18].

Europol, in its 2024 Internet Organized Crime Threat Assessment, warned of a “rapidly developing criminal ecosystem” around AI, with the potential to scale attacks beyond anything previously seen [14].

8. Regional Focus: Germany’s AI Security Landscape

Germany plays a central role in shaping both AI adoption and cybersecurity governance in Europe. With strong industrial bases, advanced digital infrastructures, and significant regulatory activity, Germany faces both opportunities and threats from AI in information security.

8.1 Regulatory and Strategic Initiatives

Germany has been especially proactive in integrating AI into its cybersecurity strategy. And that’s for a reason.

- The BSI’s annual Lagebericht zur IT-Sicherheit (2024) noted an increase in AI-enhanced phishing targeting municipal governments and small enterprises [15].
- The country report for Germany by the Alliance4Europe lists many examples of AI driven disinformation and election interference, e.g., by deepfake abuse and false postings on social media in the 2025 Bundestag elections [19].

Germany has also piloted the use of AI in its national CERT (CERT-Bund) for prioritizing vulnerability alerts and correlating data from honeypots, improving its time-to-response significantly.

8.2 Legal Landscape and the EU AI Act

The countries in Europe are bound by the upcoming EU Artificial Intelligence Act (AI Act), which was passed by the European Parliament in March 2024 and is now in phased implementation [20]:

- The AI Act classifies AI systems used in cybersecurity as high-risk if they affect critical infrastructure or involve biometric data.
- Organizations must ensure transparency, robust documentation, risk management systems, and human oversight.

Switzerland, though not in the EU, is aligning its national AI governance with the Act to maintain data flow and compatibility with EU partners. However, certain differences may occur [21].

Zum Autor:

Prof. Dr. Volker Scheidemann, Studiengangleiter Bachelor Business Information Management an der Provadis School of International Management and Technology AG, verfügt über 20 Jahre Erfahrung in der Information Security Branche. Seine fachlichen Schwerpunkte an der Provadis Hochschule umfassen Mathematik, Informationssicherheit & IT-Risikomanagement und Security Awareness.

+49 172 2093794

volker.scheidemann@provadis-hochschule.de

References

[1] Morgan Stanley (2024). AI and Cybersecurity: A New Era, Sept. 2024. URL: https://www.morganstanley.com/articles/ai-cybersecurity-new-era (last called 2025/07/08)

[2] NACD, AI as a Cybersecurity Risk and Force Multiplier, Mar. 2025, URL: https://www.nacdonline.org/all-governance/governance-resources/governance-research/director-handbooks/DH/2025/ai-in-cybersecurity/ai-as-a-cybersecurity-risk-and-force-multiplier/ (last called 2025/07/08)

[3] OpenAI et al. (2025). Threat Intelligence Report: Malicious Use of AI, June 2025. URL: https://cdn.openai.com/threat-intelligence-reports/5f73af09-a3a3-4a55-992e-069237681620/disrupting-malicious-uses-of-ai-june-2025.pdf (last called 2025/07/08)

[4] IBM Security (2025). X-Force Threat Intelligence Index 2025. URL: https://www.ibm.com/thought-leadership/institute-business-value/report/2025-threat-intelligence-index (last called 2025/07/09)

[5] IBM Security (2024). Cost of a Data Breach Report 2024, Ponemon Institute. URL: https://cdn.table.media/assets/wp-content/uploads/2024/07/30132828/Cost-of-a-Data-Breach-Report-2024.pdf (last called 2025/07/08)

[6] IBM (2025). Was ist User and Entity Behavior Analytics (UEBA)? URL: https://www.ibm.com/de-de/topics/ueba (last called 2025/07/09)

[7] Google Cloud Blog (2023). Supercharging security with generative AI. URL: https://cloud.google.com/blog/products/identity-security/rsa-google-cloud-security-ai-workbench-generative-ai (last called 2025/07/09)

[8] Abuadbba, A. et al. (2025). From Promise to Peril: Rethinking Cybersecurity Red and Blue Teaming in the Age of LLMs. URL: https://arxiv.org/html/2506.13434v1 (last called 2025/07/09)

[9] Security Bridge Blog (2025). SAP Security in the Age of AI: Shifting the Advantage from Attackers to Defenders. URL: https://securitybridge.com/blog/sap-security-ai-shifting-the-advantage/ (last called 2025/07/09)

[10] FBI (2024). Criminals Use Generative Artificial Intelligence to Facilitate Financial Fraud. URL: https://www.ic3.gov/PSA/2024/PSA241203 (last called 2025/07/09)

[11] Microsoft (2024). Into the Lion's Den. URL: https://news.microsoft.com/wp-content/uploads/prod/sites/626/2024/05/Cyber_Signals_Issue_7_May_2024-2.pdf (last called 2025/07/09)

[12] The Guardian (2024). Company worker in Hong Kong pays out £20m in deepfake video call scam. URL: https://www.theguardian.com/world/2024/feb/05/hong-kong-company-deepfake-video-conference-call-scam (last called 2025/07/08)

[13] University of Sheffield (2023). Security Risks in AIs such as ChatGPT revealed by researchers. URL: https://sheffield.ac.uk/news/security-threats-ais-such-chatgpt-revealed-researchers (last called 2025/07/08)

[14] Europol (2024). IOCTA: Internet Organized Crime Threat Assessment. URL: https://www.europol.europa.eu/publication-events/main-reports/internet-organised-crime-threat-assessment-iocta-2024 (last called 2025/07/08)

[15] BSI (2024). Lagebericht zur IT-Sicherheit in Deutschland 2024. URL: https://www.bsi.bund.de/DE/Service-Navi/Publikationen/Lagebericht/Archiv-Lageberichte/2024/lagebericht_2024_node.html (last called 2025/07/08)

[16] ZDF WiSo (2024). Was tun bei Telefonbetrug mit KI-Stimmen. URL: https://www.zdfheute.de/ratgeber/betrugsmasche-telefonbetrug-ki-deepfake-100.html (last called 2025/06/30).

[17] The Guardian (2025). AI tools used for child sexual abuse images targeted in Home Office crackdown. URL: https://www.theguardian.com/technology/2025/feb/01/ai-tools-used-for-child-sexual-abuse-images-targeted-in-home-office-crackdown (last called 2025/07/08)

[18] Australian Federal Police (2024). Victorian man jailed for producing almost 800 AI-generated child abuse images. URL: https://www.afp.gov.au/news-centre/media-release/victorian-man-jailed-producing-almost-800-ai-generated-child-abuse-images (last called 2025/07/08)

[19] Country Report: Assessment of Foreign Information Manipulation and Interference (FIMI) in the 2025 German Federal Election (2025). URL: https://alliance4europe.eu/foreign-information-manipulation-2025-german-federal-election (last called 2025/07/21)

[20] EU Artificial Intelligence Act (2024). URL: https://artificialintelligenceact.eu/ (last called 2025/07/21)

[21] New Pathway of Regulating Artificial Intelligence in Switzerland: Competitive Edge or Challenge? (2025). URL: https://www.sidley.com/en/insights/publications/2025/03/new-pathway-of-regulating-artificial-intelligence-in-switzerland-competitive-edge-or-challenge (last called 2025/07/21)