KI – ein IKT-Risiko?

KI hat zwischenzeitlich Einzug in den Bankenalltag gefunden. Grund genug, sich dem Themenfeld KI aus der Perspektive des IKT-Risikos und der Informationssicherheit anzunähern.

Die großen Sprachmodelle („large language models", LLM) wie ChatGPT haben künstliche Intelligenz (KI) nicht nur der breiten Öffentlichkeit zugänglich gemacht, sondern auch in der Genossenschaftlichen FinanzGruppe zu einem festen Bestandteil der täglichen Arbeit werden lassen. Sei es beim Erstellen von Präsentationen, Übersetzen von Texten oder zur Generierung ansprechender Grafiken: KI-Systeme zu nutzen, ist in zahlreichen Prozessen der Bank inzwischen etabliert. Deshalb soll hier das Themenfeld „KI" aus dem Blickwinkel des IKT-Risikomanagements und der Informationssicherheit betrachtet werden. Der Fokus liegt auf potenziellen und realen Bedrohungen, aber auch auf den Chancen, Ihre IKT- und Informationssicherheitsprozesse auszubauen.

Für uns gilt grundsätzlich, dass KI-gestützte Systeme weder als Universallösung für alle Anforderungen zu preisen, noch als generelle Bedrohung anzusehen sind.Sie sind Werkzeuge wie viele andere und können daher sowohl konstruktiven als auch destruktiven Verwendungszwecken dienen (sog. Dual-Use-Güter). Ein Beispiel hierzu wären Penetrationstests, die sowohl beim Red Teaming einen Angriff auf ein Zielnetzwerk simulieren können als auch bei der Durchführung tatsächlicher Angriffe Verwendung finden. Oder Sprachmodelle, die den Anwender anhand seiner Stimme – nahezu einwandfrei – identifizieren können und ihm so eine einfache Anmeldung an einem System ermöglichen. Andererseits wird mithilfe derselben Modelle auch Identitätsdiebstahl vereinfacht.

Einfluss von KI am Beispiel von Sprachmodellen

Der Start von ChatGPT im November 2022 führte zu einem intensiven Wettbewerb auf dem Markt für Chatbots. Seitdem sind zahlreiche weitere Sprachmodelle auf den Markt gekommen mit teilweise erheblichen Leistungssprüngen. Die allgegenwärtige Verfügbarkeit leistungsstarker Sprachmodelle beeinflusst seitdem nicht nur zahlreiche Branchen in ihren Kernprozessen, sondern wird aller Voraussicht nach auch den Cybersicherheitssektor nachhaltig beeinflussen.

Wir bei der DZ CompliancePartner setzen uns daher bereits seit geraumer Zeit mit Sprachmodellen auseinander. Auch ohne Kenntnisse der Sprache generieren Angreifer beispielsweise mittels Sprachmodellen qualitativ hochwertige Phishing-Nachrichten. Textbausteine können hierbei mit zusätzlichem Inhalt ergänzt werden, um die Nachrichten zu personalisieren oder einen bestimmten Schreibstil zu verwenden – was zu überzeugenden Nachrichten führt. Bisher etablierte Ansätze zur Identifizierung von betrügerischen Nachrichten, wie z. B. die Prüfung auf Rechtschreibfehler und unkonventionellen Sprachgebrauch, reichen heute nicht mehr aus. Dies sollte im Rahmen der Sensibilisierung und des Awareness-Trainings Ihrer Mitarbeitenden thematisiert werden.

Die Kombination von Sprachmodellen mit weiteren generativen KI-Techniken, wie z. B. Deepfakes (realistisch wirkende verfälschte Medieninhalte) in Webmeetings, ermöglicht es, Angriffe von zuvor unerreichter Qualität durchzuführen. So wurde vor circa einem Jahr bereits ein Mitarbeiter eines internationalen Konzerns in Hongkong im Rahmen einer Videokonferenz mittels KI-generierten Teilnehmern dazu verleitet, rund 24 Mio. Euro an Betrüger zu transferieren.[1] Sämtliche Teilnehmer des Meetings mit Ausnahme des betreffenden Angestellten bestanden aus KI generierten „Personen", die allesamt von den tatsächlich existierenden Personen nicht zu unterscheiden waren.

Weiterhin werden ChatGPT und seine Ableger auch bei der Schadcode-Generierung vermehrt genutzt. Die Einstiegshürden für Personen, die bösartige Aktivitäten durchführen wollen, sind stark gesunken, da die Generierung und Verbreitung von hochpotentem Schadcode keine besonderen Anforderungen mehr an die technische Qualifikation der Angreifer stellt.[2] Die Maßnahmen zur Verhinderung von Missbrauch erstrecken sich dagegen bisher allenfalls auf rudimentäre Filtersysteme, die allzu böswillige Prompts wie „Erstelle mir ein Script für Denial-of-Service-Angriffe" abwehren. Mit ein wenig Experimentierfreude sind derartige Maßnahmen jedoch leicht zu umgehen.

So existieren heute KI-Systeme, die Teilschritte von Cyberangriffen durchführen können: Mithilfe von ChatGPT sowie der Assistants API von OpenAI können Entwickler beispielsweise Anwendungen mit anspruchsvollem Copilot-ähnlichem Verhalten erstellen, die Daten durchsuchen, Lösungen vorschlagen und Aufgaben automatisieren. Angreifer wie auch Pentester verwenden diese Verfahren, um beispielsweise in der Aufklärungsphase eines Cyberangriffs Serverantworten automatisiert zu analysieren. Weiterhin können hiermit auch SQL Injections (Ausnutzen von Schwachstellen im Quelltext von Anwendungen, um beispielsweise Schadcode einzubinden) oder Brute-Force-Angriffe (Erraten von Passwörtern oder anderer Codes durch massenhaftes Ausprobieren) äußerst effizient durchgeführt werden.[3]

Grenzen aktueller KI-Systeme

Für umfangreiche, breite Angriffsszenarien durch technisch kaum oder wenig versierte Täter ist eine KI erforderlich, die alle sechs Schritte einer Cyberattacke autonom durchführen kann:

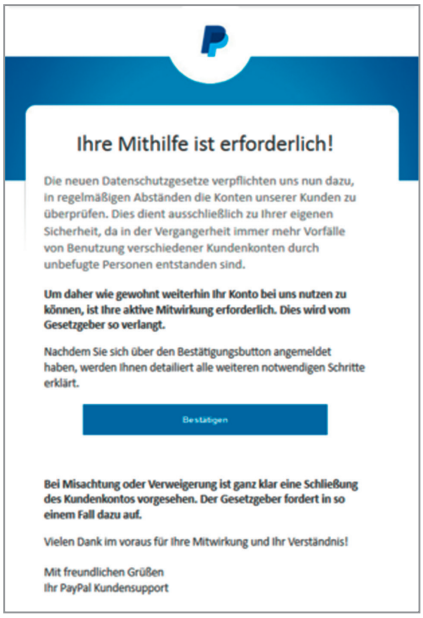
- Aufklärung,
- Rüstung und Bereitstellung,
- Ausnutzung,
- Installation,
- Erlangung von Befehls- und Kontrollgewalt sowie
- Ausführung (des Schadcodes, des Datendiebstahls etc.).

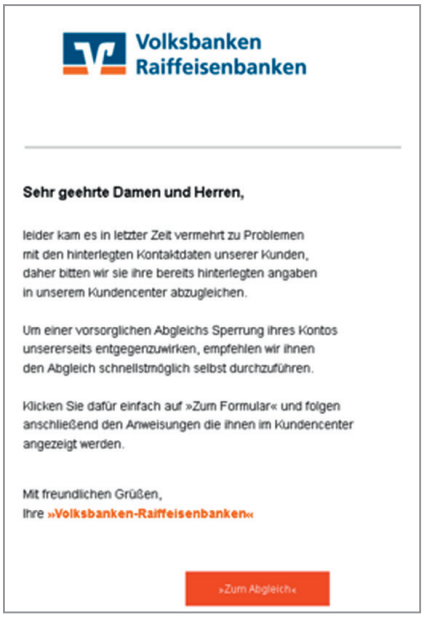
Aus der Sicht eines Penetrationstesters könnten derartige Werkzeuge selbstverständlich auch effizient dazu genutzt werden, um die Widerstandsfähigkeit von IT-Systemen zu verifizieren und somit den Zeit- und Kostenaufwand für die Durchführung von Penetrationstests zu optimieren. Obschon die Leistungsfähigkeit sowie auch die Abstraktionsfähigkeit der aktuellen KI-Systeme auf einem beeindrucken Niveau angekommen sind und fortlaufend weiter geforscht wird, existieren aktuell zumindest – noch – keine Tools für tatsächliche vollautonome Angriffsszenarien.[4]

Vor dem Hintergrund sich stetig verändernder Bedrohungslagen und damit einhergehender aufsichtsrechtlicher Anforderungen, wie z. B. DORA, ist es bedeutsam, dem Thema Cybersicherheit eine erhöhte Aufmerksamkeit einzuräumen. Ihre Fähigkeit, auf neue Bedrohungsszenarien schnell und wirksam zu reagieren bzw. diese im besten Fall vorherzusehen, wird zukünftig maßgeblich für ein hinreichendes IKT- und Informationssicherheitsniveau sein. Nur **technische und organisatorische Maßnahmen im Zusammenspiel** können dies gewährleisten. Da KI häufig klassische Angriffe verstärkt, fallen die nachfolgend dargestellten Maßnahmen weitgehend auch in den Bereich der klassischen IKT-Risikokontrolle und Informationssicherheit.

Technische Maßnahmen

- **Detektion und Reaktion auf Angriffe:**
  Mittels KI-gestützter Tools können Cyberattacken in sehr kurzer Zeit und mit hohem Präzisionsgrad lanciert werden. Eine resiliente Infrastruktur enthält Mechanismen zur zeitnahen Detektion und Reaktion (Intrusion Prevention/Detection sowie Data Loss Prevention/Detection) auf solche Angriffe, sodass Angriffe rasch erkannt und Risiken hieraus minimiert werden können. Durch die Implementierung von Automatisierungen und intelligenten (KI-gestützten) Sicherheitssystemen kann die Reaktionszeit erheblich verkürzt werden. KI-Systeme sind beispielsweise in der Lage, bei DDoS-Angriffen verdächtige **Datenströme** (beispielsweise mit einem gemeinsamen Verhaltensprofil oder aus einem ähnlichen IP-Bereich) in Echtzeit zu **analysieren,** um diese gar nicht erst zu den eigentlichen Serversystemen durchzuleiten.
- **Einsatz von Sicherheitszonen:**
  Durch die Implementierung einer **mehrschichtigen Sicherheitsarchitektur,** die sowohl physische als auch logische Sicherheitsvorkehrungen umfasst, wird die Infrastruktur besser gegen Angriffe abgesichert. Selbst wenn eine Ebene überwunden wird, bleibt der Schutz durch andere Sicherheitsmaßnahmen bestehen. Angreiferbewegungen innerhalb eines Netzwerks können mittels **Zero-Trust-Modellen** (und einer hieraus resultierenden kontinuierlichen Überprüfung und Authenti-





Angebliche E-Mails von PayPal und der Volksbanken Raiffeisenbanken. Quelle: www.verbraucherzentrale.de

fizierung aller Benutzer und Geräte, unabhängig von ihrem Standort oder ihrer Herkunft) oder **Segmentierung** eingegrenzt werden. Sehr hoch schutzbedürftige Informationen wie z. B. sensible Mitarbeiterdaten oder Geschäftsgeheimnisse können hierdurch selbst bei einer Kompromittierung eines Teilnetzes geschützt bleiben.

- **Implementierung redundanter Strukturen zur Ausfallsicherheit:**
  Der **Einsatz redundanter Strukturen** sorgt dafür, dass KI-basierte Angriffe Ihre IT-Services nicht dauerhaft außer Betrieb setzen können, wodurch die Auswirkungen des Angriffs verringert werden.
- **Einsatz von Multi-Faktor-Authentifizierung (MFA):**
  Auch mittels MFA können Social-Engineering-Angriffe in der Praxis oft wirksam abgewendet werden; insbesondere trifft dies auf **Phishing-Angriffe** zu. Selbst wenn ein Angreifer an die Zugangsdaten eines Mitarbeiters gelangt, verhindert eine MFA, dass er sich ohne den zusätzlichen Authentifizierungsfaktor Zugang verschaffen kann. Sichern Sie daher sehr hoch schutzbedürftige respektive kritische Bestandteile Ihres Informationsverbundes stets mittels MFA ab.
- **Einsatz von Anti-Viren-Software und Firewalls:**
  Der Einsatz von E-Mail-Sicherheitslösungen, die Phishing-Versuche und schadhafte Links erkennen und blockieren, ist ein weiterer wichtiger Baustein zum Schutz. Hierdurch werden E-Mails mit dolosen Inhalten schon häufig herausgefiltert, bevor sie den Posteingang des Mitarbeitenden erreichen. Der Einsatz von Firewall-Systemen und Anti-Viren-Software erscheint bei von der Atruvia AG betriebenen IT-Systemen zwar wirksam umgesetzt, jedoch offenbart sich bei der Verwendung von (ggf. GFG-fremden) Drittanbietern hier noch Optimierungspotenzial. Dem **IKT-Dienstleistermanagement** kommt somit eine besonders wichtige Rolle zu.

Organisatorische Maßnahmen

- **Wirksames Patchmanagement:**
  Ein effektives und effizientes Patchmanagement ist entscheidend, um sich vor KI-gesteuerten Angriffen zu schützen. Es trägt dazu bei, Sicherheitslücken in der Software und in IT-Systemen zeitnah zu schließen. KI-Technologien, insbesondere maschinelles Lernen und automatisierte Angriffsstrategien, können Schwachstellen in IT-Systemen schneller und gezielter ausnutzen als traditionelle Angriffsmethoden. Durch die regelmäßige Aktualisierung von Software und das Schließen von Sicherheitslücken können potenzielle Einstiegspunkte für diese Angriffe minimiert werden. Daher ist das Patchmanagement, auch von nicht durch Atruvia AG gemanagten Bestandteilen Ihres Informationsverbundes, ein zentraler Bestandteil einer umfassenden Cybersicherheitsstrategie.
- **Anpassungsfähigkeit:**
  KI und maschinelles Lernen entwickeln sich rasant weiter, sodass sich auch Angriffstechniken ständig ändern. Eine resiliente IT-Infrastruktur ist darauf ausgelegt, kontinuierlich an neue Bedrohungen und Technologien angepasst zu werden. Dies bedeutet, dass im Rahmen von **Soll-Soll-Abgleichen die Maßnahmenkataloge** an die jeweils aktuellen Bedrohungen anzupassen sind.

- **Schulung und Sensibilisierung der Mitarbeitenden/ User:**
  Regelmäßig sowie anlassbezogen sollten alle Mitarbeitenden in den gängigen **Techniken des Social Engineering geschult** werden, um verdächtige Aktivitäten sicher erkennen zu können. Hierzu zählen Verfahren wie
  - Phishing,
  - Spear-Phishing,
  - Vishing (Voice-Phishing),
  - Pretexting oder Baiting.

  Darüber hinaus sollte mittels **simulierter Angriffe** (z. B. durch Phishing-Tests) regelmäßig überprüft werden, wie Ihre Mitarbeitenden auf solche Attacken reagieren. Scheinangriffe helfen hierbei unserer Erfahrung nach maßgeblich, die Awareness zu schärfen und auch unter Stress richtig zu reagieren.
- **Klare Kommunikationsrichtlinien:**
  Risikobehaftete Geschäftsvorfälle wie beispielsweise das Ausführen von **Überweisungen** oder der **Handel von Wertpapieren** bedürfen stets einer **Verifizierung** des Berechtigten und der **Autorisierung** durch einen Berechtigten (Kunden/Bevollmächtigten). Hierzu ist es ratsam, für derartige Geschäftsvorfälle ausschließlich als **sicher geltende Kommunikationskanäle** zu verwenden. Auch unklare oder verdächtige Anfragen beispielsweise per Telefon oder E-Mail sollten stets über einen sicheren, bekannten Kanal vor Ausführung abgeklärt werden.
- **Identitäts- und Rechtemanagement:**
  Der Zugang zu kritischen IT-Systemen und sowie der Zugriff auf hoch schutzbedürftige Informationen ist zwingend auf ein Minimum zu beschränken. Dies gilt **auch für Drittsysteme** und **Webanwendungen,** die bei Dienstleistern außerhalb des Atruvia-Umfeldes betrieben werden. Mitarbeitende (eigene und die ggf. verwendeter Dienstleister) sollten nur auf die Informationen zugreifen können, die sie für die Erledigung ihrer Aufgaben tatsächlich benötigen.[5] Weiterhin sind die so definierten Sollkonzepte regelmäßigen sowie anlassbezogenen Überprüfungen zu unterziehen.

Fazit

Sowohl Cybersicherheit als auch künstliche Intelligenz unterliegen einem ständigen Wandel. Als Verantwortliche für Informationssicherheit sind wir gut beraten, die kommenden Entwicklungen eng zu beobachten: Fakt ist, dass sich die KI-Modelle und damit auch ihre Anwendbarkeit für Cyberangriffe rasant weiter entwickeln werden. Entsprechend sind auch die Abwehrmöglichkeiten massiv zu erweitern. Die Implementierung sowie der Betrieb der hierzu erforderlichen technischen Maßnahmen obliegt, bei ausschließlicher Nutzung von Atruvia-Services, in der Regel dem Rechenzentrum. Banken unserer Gruppe fällt intern insbesondere die Aufgabe zu, die zuvor dargestellten organisatorischen Maßnahmen zu implementieren und die Mitarbeitenden zu schulen. ▬

Björn Scherer
Beauftragter IKT-Risikokontrolle und Informationssicherheit
E-Mail: bjoern.scherer@dz-cp.de

[1] https://www.heise.de/news/Videokonferenz-voller-KI-Klone-Angestellter-schickt-Betruegern-24-Millionen-Euro-9618064.html, abgerufen am 28.12.2024
[2] Vgl. Van Eeten, Michel, et al. An Attacker's Dream? Exploring the Capabilities of ChatGPT for Developing Malware, Delft University of Technology, 2023.
[3] Vgl. Fang, Richard, et al. LLM Agents can Autonomously Hack Websites: https://openreview.net/pdf?id=6xubl2J2VP , abgerufen am 04.01.2024
[4] Vgl. Wenig, Lilian. LLM Powered Autonomous Agents: https://lilianweng.github.io/posts/2023-06-23-agent/, abgerufen am 04.01.2024
[5] Vgl. AT 4.3.1 Tz. 2 der MaRisk