Deepfakes verstehen und bekämpfen:

Ein Leitfaden für die iT-Sicherheit in Unternehmen

DISCOVER THE SPIR.iT OF EXCELLENCE.
SURPASS YOUR SUCCESS.

ZIELE DIESES WHITEPAPERS

Die Digitale Transformation hält viele Vorteile und Chancen für unsere Gesellschaft und Wirtschaft bereit – aber natürlich bringt auch sie neue Herausforderungen und Risiken.

Durch die Kombination von generativer Künstlicher Intelligenz, exponentiellem technischen Fortschritt und der 'Kreativität' von Cyberkriminellen sind Deepfakes in kürzester Zeit zu einer sehr ernstzunehmenden Bedrohung unserer Wirtschaft und Gesellschaft geworden – und die Bedrohungslage nimmt derzeit täglich weiter zu.

Cyberangriffe, die auf Deepfake-Technologie beruhen, heben bekannte Angriffsszenarien wie CEO-Fraud, Phishing und Social-Engineering auf ein ganz neues Bedrohungslevel. Zudem führen sie dazu, dass gerade im Unternehmenskontext plötzlich Angriffsszenarien wie Erpressung, Identitätsdiebstahl und Desinformationskampagnen einen immer wichtigeren Stellenwert einnehmen.

Auch wenn bereits an konkreten Lösungen für dieses komplexe neue Thema geforscht und gearbeitet wird: Bis zu ihrer erfolgreichen und weitreichenden Implementierung werden noch einige Jahre vergehen. In der Zwischenzeit sind und bleiben Deepfake-Angriffe ein Cyberrisiko mit – wie wir in diesem Whitepaper detaillierter ausführen werden – teilweise existenzbedrohenden Auswirkungen.

Mit diesem Whitepaper möchten wir daher vor allem eines erreichen: Möglichst viele Leser*innen für dieses wichtige Thema sensibilisieren, damit wir gemeinsam den Kriminellen sowohl in unserem beruflichen als auch privaten Alltag so wenig Angriffsfläche wie möglich bieten und so im Sinne unserer eigenen Unternehmen als auch gesamtgesellschaftlich den Schaden weitgehend minimieren. Denn der wirkungsvollste Schutz vor Deepfake-basierten Angriffsversuchen von Cyberkriminellen besteht darin, dass möglichst viele Menschen wissen, dass es diese neue Art der Medienmanipulation gibt, was sich hinter Deepfakes verbirgt und wie sie funktionieren – dann ist das Risiko, dass wir ihr zum Opfer fallen, schon erheblich reduziert!

DIE RAHMENBEDINGUNGEN

DIGITALE TRANSFORMATION

Bei R.iT sensibilisieren wir unsere Kunden in Beratungsgesprächen und Workshops immer für den exponentiellen Charakter der Digitalen

Transformation. Denn aufgrund der Tatsache, dass der Treiber hinter der Digitalen Transformation moderne Informationstechnologie ist und sich deren Leistungsfähigkeit schon seit Jahr-zehnten exponentiell entwickelt, sind wir mittlerweile in einem Stadium angekommen, in dem die Leistungssteigerungen pro Zeiteinheit (und folglich die Auswirkungen der damit zusammenhängenden Innovationen pro Zeiteinheit) spürbar zunehmen und dazu führen, dass sich die Rahmenbedingungen, unter denen wir Entscheidungen treffen, immer schneller verändern (Grafik s. u.)

Die größte Herausforderung dabei ist, dass exponentielle Entwicklungen für uns Menschen ein nur extrem schwer zu fassendes Phänomen sind, denn unser Gehirn kann nicht mit exponentiellen Zusammenhängen umgehen – wir kennen uns nur mit linearen aus.

Die Erwartung, dass sich die Zukunft 'jedenfalls wenigstens so ungefähr' entwickeln wird, wie die Vergangenheit, ist der gefährlichste Trugschluss, dem wir in der heutigen Zeit sowohl im unternehmerischen als auch im privaten Umfeld unterliegen können. Während wir die Gründe und Auswirkungen in einem eigenen Whitepaper näher ausgeführt haben[1], werden wir uns an dieser Stelle darauf fokussieren, welche Konsequenzen die exponentiellen Leistungssteigerungen moderner Informationstechnologie auf den Bereich der iT-Sicherheit haben.

Künstliche Intelligenz

Das wohl prominenteste Beispiel für das, was die exponentielle Leistungssteigerung im Bereich der iT auf unser aller Leben an Auswirkungen hat, ist Künstliche Intelligenz. Während Anfang November 2022 außer KI-Experten wohl kaum jemand das Thema auf der Agenda hatte, änderte sich das schlagartig mit der Erscheinung von ChatGPT 3.5 am 30. November 2022: Innerhalb von nur fünf (!) Tagen hatten sich mehr als 1.000.000 Nutzer registriert; und im Januar 2023 – also nach etwa sechs Wochen – waren es schon 100.000.000 Nutzer.

Mindestens ebenso spannend ist die exponentielle Entwicklungsgeschwindigkeit der technologischen Basis für generative Künstliche Intelligenz: Während GPT-1 (die erste Version) 2017/18 aus 'nur' 117 Mio. vortrainierten Parametern bestand, kam GPT-2 im Februar 2019 bereits auf 1,5 Mrd.; nur weitere 15 Monate später erhöhte sich diese Zahl bei GPT-3 schon auf 175 Mrd. Parameter.[2]

Was sich hier sehr technisch anhört, hatte zur Folge, dass aus einem zunächst sehr eingeschränkt nutzbaren Tool ein allgemein einsetzbares Werkzeug wurde, das plötzlich qualitativ hochwertige Texte in unterschiedlichsten Sprachen schreiben, Zusammenhänge erklären, Besprechungen und Texte zusammenfassen und später dann sogar Bilder erstellen und interpretieren konnte.

Nicht zuletzt aufgrund des damit ausgelösten 'Hypes' und des in KI-Unternehmen investierten Kapitals wurden ChatGPT und vergleichbare KI-basierte Anwendungen innerhalb von 1,5 Jahren damit zu omnipräsenten Werkzeugen.

[1] vgl. Rademann, Tobias (2019): „Digitalisierung: Auf der zweiten Hälfte des Schachbretts"; R.iT-Digitalisierungs-Whitepaper Nr. 2; https://www.rit.de/media/pages/digitale-transformation/aa1aa71a99-1628613442/r.it_-_digitalisierungswhitepaper_nr._2.pdf

[2] Quelle: https://de.wikipedia.org/wiki/ChatGPT

*Innovationen im iT-Umfeld als **Treiber** für alle anderen Bereiche*

Deepfakes: Anwendungsbereiche

DEEPFAKES – EINFÜHRUNG

Die exponentielle Leistungssteigerung im Bereich der generativen Künstlichen Intelligenz hat in kürzester Zeit dazu geführt, dass digitale Inhalte nicht nur erheblich leichter, schneller und kostengünstiger erstellt, sondern eben auch verändert (negativ ausgedrückt: manipuliert) werden können.

Sofern (generative) Künstliche Intelligenz dazu genutzt wird, manipulierte Text-, Bild-, Audio- oder Videodateien zu erstellen, spricht man von 'Deepfakes'. Das Ziel von Deepfakes besteht also darin, mediale Inhalte zu schaffen, bei denen weder Menschen noch Systeme in der Lage sind, zu erkennen, dass es sich hier um manipulierte Inhalte handelt.

Der Begriff 'Deepfake' ist eine Zusammensetzung aus den englischen Worten 'deep learning' und 'fake'. 'Deep learning' ist eine Methode des maschinellen Lernens, die auf künstlichen neuronalen Netzen basiert, die große Mengen an Daten verarbeiten können. 'Fake' bedeutet gefälscht oder unecht. Deepfake bezeichnet also eine Technik, die mit Hilfe von deep learning/ generativer KI digitale Inhalte wie Texte, Audios, Bilder oder sogar Videos manipuliert oder erzeugt, die zwar nicht der Realität entsprechen, aber diese täuschend echt nachahmen.[3]

Die Einsatzbereiche für Deepfakes sind dabei weiträumig: Sie reichen von innovativen Marketingstrategien mit personalisierter Werbung über die Optimierung des Kundenservice (bspw. durch den Einsatz realistischer Avatare) sowie die Restaurierung und Verbesserung medialer Inhalte (bspw. alter Filme oder Fotos) aber auch bis hin zu kriminellen Aktivitäten (s. u.).

Trotz der hohen Zahl positiver möglicher Anwendungsbereiche werden Deepfakes derzeit 'leider' noch primär mit kriminellen Aktivitäten verbunden; wie wir nachfolgend detailliert ausführen werden, missbrauchen Kriminelle dieses neue Werkzeug gerade massiv für ihre Zwecke. Es bleibt daher zu hoffen, dass diese durchaus sehr vielversprechende Technologie in Zukunft mehr und mehr für nutzbringende Anwendungen eingesetzt wird.

Bei Deepfakes geht es sowohl um die Manipulation von personenbezogenen Inhalten (Stimme, Fotos, Videosequenzen) als auch um die von Ereignissen (Fotos, Videosequenzen). Bekannte Beispiele sind:

- **Deepfakes mit personenbezogenen Inhalten**
  - Radiowerbung der BILD-Zeitung mit der Stimme des Bundeskanzlers[4]
  - Werbevideo der BILD-Zeitung mit der Stimme und der Person des Bundeskanzlers[5]

[3] Für weitergehende Informationen siehe auch die Seite des BSI zum Thema „Deepfakes – Gefahren und Gegenmaßnahmen"; https://www.bsi.bund.de/DE/Themen/Unternehmen-und-Organisationen/Informationen-und-Empfehlungen/Kuenstliche-Intelligenz/Deepfakes/deepfakes_node.html

[4] Audiosequenz bspw. unter: https://www.ndr.de/nachrichten/info/Deepfakes-in-der-Werbung-Scholz-wirbt-angeblich-fuer-die-Bild,audio1589816.html

[5] Videosequenz unter https://www.google.com/search?client=firefox-b-d&sca_esv=fd72789be247385b&sxsrf=ADLYWIJbxrKalCyDEXXlJsHKIhQxh8xG8w:1721287786128&q=BILD+Werbung+Scholz+KI+YouTube&sa=X&ved=2ahUKEwjmn7HYiLCHAxXEAtsEHaYYCQkQ1QJ6BAhFEAE&biw=2361&bih=1377&dpr=1#fpstate=ive&vld=cid:c9b2ffc8,vid:A0FoXO3DcN8,st:0

- Videosequenzen mit Gesichtern und Stimmen von Prominenten (wie Christian Sievers), die für extrem lukrative Investments werben[6]

- **Deepfakes mit ereignisbezogenen Inhalten**
  - Bombenanschlag auf das Pentagon[7]
  - Argumente von Musks Anwälten, vorgelegte Beweise der Ankläger seien Deepfakes gewesen[8]

Prinzipiell ist die Manipulation von medialen Inhalten/Identitäten nichts Neues; allerdings war diese in den vergangenen Jahr(zehnt)en enorm ressourcen-, kosten- und zeitintensiv, sofern man annähernd akzeptable Qualität erhalten wollte. Die eingangs beschriebenen Fortschritte im Bereich der generativen KI im Zusammenhang mit den exponentiellen Leistungssteigerungen moderner Hard- und Software führen nun jedoch dazu, dass es schrittweise erheblich leichter, günstiger und schneller möglich wird, authentisch erscheinende, aber manipulierte Texte, Fotos sowie Audio- und Videosequenzen in immer besserer Qualität zu erstellen. Zugegebenermaßen: Noch ist die Erstellung täuschend echter Deepfakes v. a. im Videobereich ziemlich kostspielig; aber auch hier wird der technische Fortschritt in sehr naher Zukunft dazu führen, dass diese letzte kostenintensive Bastion bald fällt.

DEEPFAKES ALS SICHERHEITSRISIKO

Wie oben bereits angedeutet, lassen sich natürlich auch Deepfakes für kriminelle Zwecke einsetzen – und Kriminelle investieren gerade sehr viel Zeit, Geld und Kreativität in die Erstellung ausgefeilter und sehr wirkungsvoller Deepfake Angriffsszenarien.

So nutzen sie die erzeugten falschen bzw. manipulierten Medieninhalte bspw. für folgende Zwecke:

- Desinformation und politische Propaganda
- Rufschädigung
- Erpressung
- Informationsdiebstahl/Social Engineering
- Identitätsdiebstahl
- Betrug

Bei Deepfakes mit personenbezogenen Manipulationen lassen sich die folgenden Szenarien unterscheiden[9]:

Fälschung von Stimmen

- **Text to Speech**

  Bei diesem Anwendungsszenario geht es darum, mit Hilfe generativer KI aus einem geschriebenen Text eine Audio-Sequenz zu erstellen, die der Stimme einer realen Person entspricht und weder für Menschen noch für Systeme von der Originalstimme unterscheidbar ist.

  So lassen sich Aufnahmen erstellen (und verbreiten), in denen bekannte Personen Aussagen treffen, die sie so in der Realität nie getroffen haben, bspw. zu kontroversen Themen, zur wirtschaftlichen Situation ihres Unternehmens, zur Verstrickung in illegale Aktivitäten, etc.

  Diese Form von Deepfakes wird u. a. zu Erpressungsversuchen, für Betrug und zur Desinformation genutzt (s. u.).

- **Voice Conversion**

  Bei der Konvertierung von Stimmen wird die Stimme des Angreifers in die eines anderen (die Stimme des Opfers oder die Stimme, die einer Person, die dem Opfer bekannt ist) konvertiert.

  Hierdurch lassen sich bspw. in realen Telefongesprächen in Echtzeit Stimmen imitieren und Inhalte vermitteln, die die betreffende Person nie gesagt hat.

  Angriffsszenarien, in denen diese Methode verwendet wird, sind u. a. CEO-Fraud (s. u.) oder im privaten Umfeld die moderne Variante des Enkeltricks.

[6] vgl. bspw. ein Deepfake von Christian Sievers unter https://x.com/CHSievers/status/1704505532571152657 oder weitere Beispiele unter https://www.onlinesicherheit.gv.at/Services/News/Deepfake-Videos-mit-bekannten-Gesichtern-locken-in-Investmentfallen.html [7] bspw. https://www.spiegel.de/netzwelt/netzpolitik/pentagon-fake-bild-von-explosion-sorgt-fuer-aufregung-a-d4510a09-07d6-4d63-b72d-2c3bf28b52a9

[7] bspw. https://www.spiegel.de/netzwelt/netzpolitik/pentagon-fake-bild-von-explosion-sorgt-fuer-aufregung-a-d4510a09-07d6-4d63-b72d-2c3bf28b52a9

[8] siehe bspw. https://www.derstandard.de/story/2000145971564/richter-schmettert-argument-der-tesla-anwaelte-ab-musk-zitate-seien

[9] zur Übersicht siehe: BSI (2024) „Deepfakes – Gefahren und Gegenmaßnahmen"; https://www.bsi.bund.de/DE/Themen/Unternehmen-und-Organisationen/Informationen-und-Empfehlungen/Kuenstliche-Intelligenz/Deepfakes/deepfakes_node.html

KI oder real – das ist hier die Frage!

Fälschung von Gesichtern[10]

- **Gesichtstausch (face swapping)**
  Beim Face Swapping wird das Gesicht eines Angreifers in das des Opfers 'eingefügt' bzw. 'kopiert'. Auf diese Art lassen sich bspw. Mimik, Kopfbewegungen, Blickrichtung und Beleuchtungssituation eines Angreifers auf den Kopf des Opfers projizieren. Damit können Videosequenzen (teilweise mittlerweile sogar in Echtzeit) erzeugt werden, die zwar das Opfer zeigen, hinter denen aber tatsächlich ein Angreifer steckt.

  Diese Art der Deepfakes kann – natürlich in Kombination mit Voice Conversion – für eine breite Palette an Deepfake-Angriffen verwendet werden, u. a. für CEO-Fraud, Erpressung, Phishing/Social-Engineering.

- **Manipulation des Gesichts (Face Reenactment)**
  Eine weitere Art, Deepfakes für kriminelle Angriffe zu erstellen, besteht in der Manipulation von Kopf- und Lippenbewegungen sowie der Mimik: Diese werden (i. d. R. auf Basis von 3D-Modellen des Angreifers und des Opfers) von Kameraaufnahmen des Angreifers mit vorhandenen Videodaten des Opfers kombiniert. Auf diese Art lassen sich täuschend echte Videos erstellen. Auch in diesem Bereich bringt der Einsatz von Deepfake-Technologie eine ganz neue Dynamik in einen schon länger erforschten Bereich.[11]

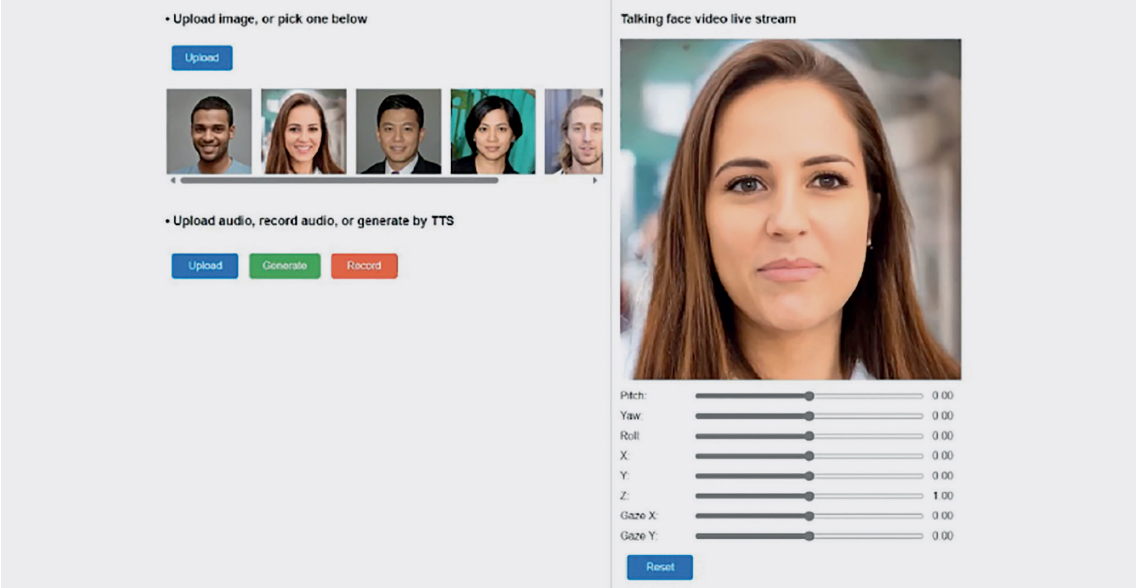
  Die Anwendungsbereiche gleichen denen des Face Swappings (s. o.).

- **Schaffung neuer Identitäten**
  Die Schaffung neuer Identitäten durch die Synthetisierung von unterschiedlichen Gesichtern stellt das letzte Deepfake-Szenario dar. Da es sich hierbei nicht um die Manipulation vorhandener Personen handelt, mag man auf den ersten Blick meinen, dass dieser Ansatz im Bereich der Cyberkriminalität eher von untergeordneter Relevanz ist.

  Dies ist jedoch ein Trugschluss: Denn die so erschaffenen künstlichen Fotos können von Cyberkriminellen bspw. im Rahmen von (falschen) Social-Media-Konten genutzt werden, um auf die jeweilige (anzugreifende) Ziel-

[10] Ein wirklich beeindruckendes, wenn auch beängstigendes Beispiel, das zeigt, was bereits heute im Bereich des Fälschens von Videos – in diesem Fall sogar basierend auf einem einzigen Foto der Zielperson! – möglich ist, kann unter https://www.microsoft.com/en-us/research/project/vasa-1/ betrachtet werden. Das Erfreuliche an diesem Projekt ist, dass es sich hier nur um ein For-schungsprojekt von Microsoft handelt und das zugrundeliegende Tool nicht öffentlich verfügbar ist; das vielleicht für viele Erschreckende ist aber, dass solche Anwendungen schon heute existieren und dass es für Cyberkriminelle vermutlich nur eine Frage des Geldes und der Zeit ist, an diese heranzukommen und sie für ihre Zwecke zu missbrauchen.

[11] Das Face Reenactment ist nicht neu. Für spannende Beispiele, die zeigen, wie weit dieser An-satz bereits 2016 ohne den Einsatz von Deepfakes war, siehe bspw. das Video von Matthias Niesser et al.: https://youtu.be/ohmajJTcpNk



VASA-1 – realistische, in Echtzeit generierte, audiogesteuerte sprechende Gesichter, Quelle: Microsoft

gruppe maßgeschneiderte, vertrauenerweckende Personas zu erstellen. Diese können dann u. a. in Social-Engineering-/Phishing-Kampagnen (s. u.) genutzt werden, um von Mitarbeiter*innen aus Unternehmen für einen Angriff relevante Informationen zu gewinnen.[12]

Neben diesen personenbezogenen Deepfakes werden ereignisbezogene Deepfakes für eine breite Palette an Angriffsszenarien wie Desinformation, Panik, Erpressung, Rufschädigung, etc. eingesetzt.

Wie anhand der o. g. Beispiele deutlich wird, eröffnen sich für Kriminelle durch die gezielte Nutzung von Deepfake-Technologien in der Tat ungeahnte neue Möglichkeiten – und für Unternehmen entstehen, wie wir in den nachfolgenden Abschnitten darlegen werden, ungeahnte neue, teilweise existenzbedrohende Risiken, auf die sie vorbereitet sein müssen und vor denen sie sich schützen sollten.

DEEPFAKE-BASIERTE ANGRIFFSSZENARIEN

Für Unternehmen sind vor allem folgende Angriffsszenarien von Relevanz, die wir in den nachfolgenden Abschnitten näher beschreiben werden:

1. Desinformation
2. Rufschädigung
3. Erpressung
4. CEO-Fraud
5. Phishing/Social Engineering
5. Identitätsdiebstahl
6. Betrug

Deepfake-Angriffstyp 1: Desinformation

Bei Desinformations-Angriffen geht es v. a. darum, mit Hilfe von Deepfake-Technologien manipulierte Medieninhalte (Texte, Fotos, Audio- und Videosequenzen) zu generieren, diese möglichst schnell, bspw. über die sozialen Medien, zu verbreiten und so die **öffentliche Meinung** zu beeinflussen.

Diese Beeinflussung kann sich auf verschiedenen Ebenen abspielen:

- Manipulation von Wahlen oder politischen Debatten
- Beeinflussung von Gerichtsverfahren oder Ermittlungen
- (kurzfristige) Marktmanipulationen
- Verursachung von Panik oder Angst in der Bevölkerung
- Spaltung oder Destabilisierung der Gesellschaft durch das Anheizen ethnischer, religiöser oder ideologischer Konflikte

[12] Einer der wohl prominentesten Fälle aus der vor-Deepfake Zeit, der eindrücklich zeigt, wie schnell sich Menschen durch falsche (und sogar nicht sehr professionell gemachte!) Social Media Profile täuschen lassen, ist die (fiktive und 2009 erschaffene) iT-Sicherheitsexpertin Robin Sage (siehe https://en.wikipedia.org/wiki/Robin_Sage).

Rufschädigung

Auswirkungen
Im Unternehmenskontext spielen Desinformationskampagnen eine eher untergeordnete Rolle, da sie sich vorwiegend auf die Manipulation der öffentlichen Meinung beziehen. Allerdings kann eine Marktmanipulation infolge einer Desinformationskampagne mindestens kurzfristig konkrete Auswirkungen auf Unternehmen haben, und auch die aus einer Panik entstehenden wirtschaftlichen Turbulenzen könnten Unternehmen treffen. Gesamtwirtschaftlich betrachtet stellen Desinformationskampagnen jedoch naturgemäß ein nicht zu unterschätzendes Risiko dar, weil sie potenziell eine größere Anzahl von Menschen erreichen und soziale, politische oder wirtschaftliche Turbulenzen verursachen.

Deepfake-Angriffstyp 2: Rufschädigung

Im Gegensatz zu Desinformationskampagnen zielen Rufschädigungs-Angriffe darauf ab, das **Ansehen oder die Glaubwürdigkeit einer bestimmten Person, Organisation oder Gruppe** zu schädigen, indem sie diese ebenfalls durch den Einsatz von Deepfakes in einem negativen oder kompromittierenden Licht darstellen.

Im Unternehmenskontext sind bspw. gedeepfakte Medieninhalte zu folgenden Themen vorstellbar:

- Falschaussagen hochrangiger Unternehmensvertreter*innen oder bekannter Persönlichkeiten zu Produkten/Dienstleistungen/dem wirtschaftlichen Status des Unternehmens/möglichen Straftaten oder Verfehlungen
- Fotos oder Videosequenzen von Personen in einer kompromittierenden oder kontroversen Situation (bspw. polarisierende politische, ethische oder religiöse Aussagen, pornografische Situationen, etc.)
- Fotos oder Videosequenzen zu einem Ereignis, das nicht existiert oder anders abgelaufen ist (bspw. Kakerlaken in einem Restaurant, Ratten in einem Lagerhaus, Kinderarbeit in Fabriken, etc.)
- Fotos oder Videos als falsche Beweise in (Gerichts-)Verfahren
- die schlichte Behauptung, vorgelegte Beweise wie Fotos, Videos, etc. seien Deepfakes, um die Glaubwürdigkeit dieser Beweise zu erschüttern bzw. Zweifel zu säen.

Auswirkungen
Auf den ersten Blick mögen die o. g. Beispiele v. a. in einem Unternehmenskontext zunächst nicht wirklich kritisch erscheinen, zumal man solchen Rufschädigungskampagnen natürlich auch medial begegnen kann, wenn man gut auf einen solchen Angriff vorbereitet ist. Multipliziert man aber die Leichtigkeit, mit der solche grundlegenden Deepfakes (v. a. Texte und Fotos) schon heute generiert werden können, mit der Geschwindigkeit, in der sich Informationen über die sozialen Medien verbreiten (lassen), dann wird schnell klar, dass Rufschädigungskampagnen eben nicht nur im politischen Bereich von hoher Relevanz sind, sondern sehr wohl mindestens kurzfristig weitreichende Konsequenzen für Unternehmen haben können, v. a. wenn diese im Consumer-Bereich tätig sind.

Sofern das betroffene Unternehmen sich proaktiv auf das Eintreten eines solchen Angriffs vorbereitet hat und über einen entsprechenden Notfall- und Kommunikationsplan verfügt, wird es in der Lage sein, das Schadensausmaß erheblich zu begrenzen. Dennoch werden offizielle Stellungnahmen, eine Analyse der gefälschten Inhalte durch Experten, Pressekonferenzen und möglicherweise der Einbezug von Cybersicherheits-

experten und der Strafverfolgung durchaus für einen merklichen Zeitraum die Ressourcen des Unternehmens binden und es somit davon abhalten, dem Tagesgeschäft wie gewohnt nachzugehen und/oder sich mit anderen wichtigen, strategischen Themen wie geplant zu beschäftigen.

Sollte das betroffene Unternehmen nicht hinreichend auf das Eintreten eines solchen Angriffs vorbereitet sein, dann wird der Schaden erheblich höher ausfallen und das Unternehmen über einen längeren Zeitraum beeinträchtigen.

Deepfake-Angriffstyp 3: Erpressung

Üblicherweise ergibt sich für Cyberkriminelle kein direkter Nutzen aus Rufschädigungskampagnen, jedenfalls nicht, wenn sie für diese Art des Angriffs nicht von einem Dritten beauftragt und dafür entsprechend bezahlt werden.

Daher stellen Erpressungsangriffe eine wichtige Alternative bzw. Vorstufe der Rufschädigung dar. Diese Form des Angriffs ist gerade im Unternehmenskontext erheblich relevanter und – wie wir unten detaillierter ausführen werden – jedenfalls in der nahen Zukunft mit fast unabsehbaren Risiken verbunden.

Im Rahmen von Erpressungsangriffen werden einzelne Individuen oder Unternehmen damit unter Druck gesetzt, dass die Kriminellen androhen, sie würden gedeepfakte Inhalte (v. a. Fotos, Audio- und Videosequenzen, aber auch Texte) veröffentlichen und so der Person bzw. dem Unternehmen schaden.

Wie bereits im vorangehenden Abschnitt zu Rufschädigungskampagnen ausgeführt, sind hierbei primär zwei Ansätze denkbar:

Zum einen könnten Kriminelle Medieninhalte erstellen, die eine **Person** in einer kompromittierenden oder kontroversen Situation zeigen, und diese als Erpressungsgrundlage nutzen.

Während in einem privaten Kontext hier eher die Erpressung von Geldbeträgen im Vordergrund steht, versuchen Kriminelle im Unternehmenskontext durch diese Masche, vor allem an vertrauliche Informationen und Zugänge zu gelangen.

Die zweite Art von Erpressungs-Angriffen durch Deepfakes besteht darin, dass Kriminelle manipulierte Medieninhalte erstellen, die ein Ereignis zeigen, das nicht existiert oder anders abgelaufen ist und so versuchen, **Unternehmen** zu erpressen, indem sie androhen, Rufschädigungskampagnen zu starten, sofern das Opfer nicht bereit ist, die Forderungen zu erfüllen (s. o.).

In der breiten Mehrheit der Fälle dürfte die Erpressung von Geldbeträgen bei dieser Art des Angriffs im Vordergrund stehen, eher selten dürfte eine Verhaltensänderung o.ä. auf Seiten des Unternehmens gefordert werden.

Auswirkungen

Bei den Auswirkungen von Deepfake-basierten Erpressungsversuchen muss u. E. unterschieden werden in solche, die auf einzelne Personen abzielen und solche, die gegen Unternehmen im Ganzen gerichtet sind:

Wenn Kriminelle bislang an vertrauliche Informationen und/oder Zugänge kommen wollten, mussten sie entweder mit Social-Engineering- und/oder Phishing-Angriffen erfolgreich sein, ihre Opfer sehr lange überwachen oder technische Sicherheitslücken finden und ausnutzen. Im Vergleich dazu sind Erpressungsversuche via manipulierter Deepfake-Angriffe um ein Vielfaches (!) leichter und schneller umzusetzen – im Ergebnis können Kriminelle auf dieser Basis Mitarbeiter*innen viel schneller und problemloser dazu bewegen, ihnen diese Informationen oder Zugänge zu liefern.

Für Entscheider*innen und Cybersicherheitsverantwortliche ist es extrem wichtig zu erkennen, dass diese Form der Erpressung von Mitarbeiter*innen aus zwei Gründen ein erhebliches Risiko mit derzeit noch völlig unabsehbaren Folgen für Unternehmen darstellt: Zum einen ist die mögliche Zahl derer, die erpresst werden können, extrem groß; damit wird sich früher oder später ein Opfer finden, das bereit ist, das zu tun, was die Erpresser von ihm oder ihr verlangen.

Zum anderen ist die Weitergabe vertraulicher Informationen und/oder Zugänge bspw. zu (digitalen wie analogen!) Unternehmensressourcen ein Sachverhalt, der nicht von technischen Schutzmaßnahmen identifiziert, geschweige denn verhindert werden kann. Selbstverständlich können innerhalb der Unternehmens-iT Data-Loss

Prevention-Lösungen eingesetzt werden, um bspw. die Weitergabe von Kreditkarteninformationen, Benutzerkennungen, etc. über Teams und vergleichbare Anwendungen zu verhindern. Wenn aber ein*e Benutzer*in erpresst wird und diese Informationen über ihr privates E-Mail-Konto, über WhatsApp oder telefonisch weitergeben soll, dann lassen sich diese Kanäle nicht vom Unternehmen überwachen.

Und wenn dann ein unbefugter Zugriff auf Basis dieser Informationen erfolgt, bspw. durch die Eingabe korrekter Authentifizierungsinformationen wie Benutzername und Passwort oder durch die Nutzung eines Schlüssels, erscheint dieser Zugriff auf die geschützten Ressourcen ja als absolut legitim.

Unternehmen sind also bereits heute in einer Form angreifbar, die bislang nicht ansatzweise denkbar/realisierbar war – und diese Situation wird sich in den nächsten Monaten erheblich verschärfen. Der sich hieraus ergebende Schaden kann u. E. derzeit kaum abgeschätzt werden; zudem dürfte die Dunkelziffer auch in Zukunft extrem hoch bleiben, weil die erfolgreich erpressten Personen kaum zugeben werden, dass sie zum einen erfolgreich erpresst wurden und darüber hinaus noch sensible Informationen und/oder Zugänge an Kriminelle weitergegeben haben.

Etwas anders sieht es u. E. in Bezug auf Erpressungsversuche aus, die auf Unternehmen abzielen: Diese Erpressungsansätze bergen vermutlich nicht ansatzweise ein solches Bedrohungspotential wie die Erpressung einzelner Personen, denn eine Veröffentlichung und Verbreitung gefälschter Aussagen oder Ereignisse kann durch geeignete Kommunikationsstrategien vglsw. (!) leicht widerlegt werden.

CEO-Fraud

Dennoch wird – wie oben bereits ausführlicher beschrieben – allein die Reaktion einiges an Ressourcen binden und zudem kann ein mind. kurzfristiger Imageschaden v. a. im Konsumentenumfeld durch gesäte Zweifel zurückbleiben.

Insgesamt ist abzusehen, dass die Auswirkungen durch Deepfake-Erpressung in den nächsten Monaten massiv zunehmen und dann in den kommenden Jahren schrittweise abnehmen werden: Denn, während derzeit noch ausgereifte Technologien fehlen, die dabei helfen, manipulierte Inhalte zu erkennen bzw. die die Authentizität echter Medieninhalte (bspw. aufgrund von Wasserzeichen, kryptografischer Verfahren, etc.) bescheinigen, wird der technische Fortschritt auch hier schnell hilfreiche Lösungen bringen. Aber solange diese Technologien noch kein Standard sind und solange noch kein wirklich breites öffentliches Verständnis für die Existenz manipulierter Medieninhalte durch Deepfake-Technologien existiert, solange wer-den die Betroffenen Angst haben und mit überdurchschnittlich hoher Wahrscheinlichkeit den Erpressungsversuchen nachgeben.

Deepfake-Angriffstyp 4: CEO-Fraud

Angriffe vom Typ CEO-Fraud existieren schon seit vielen Jahren und werden von Kriminellen immer wieder sehr erfolgreich in der Praxis umgesetzt. Die durch Deepfakes zur Verfügung stehenden Möglichkeiten heben CEO-Fraud-Angriffe nun allerdings auf ein ganz neues Level:

Während klassische CEO-Fraud-Angriffe primär auf E-Mail-basierte Kommunikation, viel Druck und den Einbezug **unbekannter, externer Partner** (wie vermeintlichen Anwälten, Treuhändern, Bank- und BaFin Mitarbeiter*innen, etc.) ausgelegt waren, dreht sich der Fokus nun auf den Einbezug manipulierter Audio- und Videosequenzen von **bekannten und entsprechend befugten Entscheider*innen** wie dem Vorstand, der Geschäftsleitung oder leitenden Angestellten.

Wie in der Einleitung zu Deepfake-basierten Angriffsszenarien beschrieben, ist es schon heute problemlos und mit extrem geringem Aufwand möglich, Stimmen durch den Einsatz von Deepfake-Technologie täuschend echt zu imitieren. Dadurch lassen sich bspw. Telefonate in Echtzeit führen, in denen die*der Angerufene tatsächlich davon überzeugt ist, mit dem Vorstand oder der Geschäftsleitung zu sprechen. Werden diese Anrufe dann noch auf Basis von Mobilfunk-Technologie (und ggf. sogar noch aus Übersee) geführt, lassen sich selbst Qualitätsmängel problemlos kaschieren und werden von allen Beteiligten als 'normal' hingenommen.

Die derzeit höchste Stufe an Deepfake CEO-Fraud-Angriffen besteht darin, dass Live-Videosequenzen (bspw. in Teams- oder Zoom-Konferenzen) durch Deepfake-Technologie so manipuliert werden, dass die anderen Teilnehmenden tatsächlich glauben, sie sprächen mit dem Vorstand oder der Geschäftsleitung, obwohl sich hinter deren Videostream tatsächlich Kriminelle verbergen.

Zugegebenermaßen ist die Qualität solcher Echtzeit-Deepfakes derzeit noch vglsw. schlecht oder aber die involvierten Kosten sind noch hoch; aber aufgrund des exponentiellen Charakters der technologischen Entwicklung ist davon auszugehen, dass es bereits in wenigen Monaten dazu kommen wird, dass das Einspielen manipulierter Videos in Echtzeit keine wirkliche Herausforderung mehr darstellen wird. Spätestens dann wird das Thema Deepfake-basierter CEO-Fraud-Angriffe ganz oben auf der Agenda vieler Unternehmen stehen.

Auswirkungen

Wie schon bei den klassischen CEO-Fraud-Angriffen besteht auch bei Deepfake-CEO-Fraud-Angriffen das Ziel darin, eine*n Mitarbeiter*in eines Unternehmens dazu zu bewegen, möglichst hohe Summen auf das Konto eines Angreifers zu überweisen.

Die Auswirkungen von Deepfake-CEO-Fraud-Angriffen sind daher identisch mit denen, die wir bereits in unserem Buch 'iT-Sicherheit & Rating' beschrieben haben: Die Liquidität des betroffenen Unternehmens kann kurz- bis mittelfristig in Mitleidenschaft gezogen werden, zudem kann es zu Imageschäden und/oder zu einer Verschlechterung der Bonität kommen.

Deepfake-Angriffstyp 5: Phishing/ Social-Engineering

Bei Phishing- bzw. Social-Engineering-Angriffen auf Basis von Deepfakes verhält es sich genauso wie bei den CEO-Fraud-Angriffen: Beide Angriffsformen existierten schon lange vor dem Aufkommen von Deepfakes, aber durch deren Einsatz werden sie auf ein ganz neues (Bedrohungs-)Level gehoben und erhalten damit einen neuen Stellenwert für die Cybersicherheitsstrategien von Unternehmen.

Beide Angriffsformen laufen häufig über Telefon oder E-Mail, aber auch über andere Kanäle wie die sozialen Medien oder Video-Chat ab.

Der Unterschied zwischen Phishing- und Social-Engineering-Angriffen liegt darin, dass Phishing eine Form von Social-Engineering ist, die auf den Massenversand von betrügerischen Nachrichten an potentielle Opfer abzielt, während Spear-Phishing und Social-Engineering für verschiedene Methoden der Manipulation und Täuschung von Menschen stehen, die oft individuell zugeschnitten sind.

Das Ziel beider Angriffstypen besteht darin, das Vertrauen der Opfer zu gewinnen, ihre Emotionen auszunutzen oder ihre Neugier zu wecken, um sie dazu zu bringen, etwas zu tun, was sie normalerweise nicht tun würden: bspw. vertrauliche Informationen wie Zugangsdaten o. ä. preisgeben, Geld überweisen, Schadsoftware herunterladen oder andere unerwünschte Handlungen

ausführen. Oft nutzen die Kriminellen dabei die Unkenntnis von Sicherheitsrisiken, die Tendenz zur Nachahmung oder die Bereitschaft zur Kooperation aus.

Bei der Beurteilung von Phishing- bzw. Social-Engineering-Angriffen ist es wichtig zu verstehen, dass diese Form der Angriffe i.d.R. nur die Vorstufe für größere und komplexere Angriffe sind: Oft geht es im Rahmen dieser Angriffe zunächst darum, eine Hintertür in ein Unternehmen zu bekommen, um anschließend über Spionage/Datendiebstahl, Verschlüsselungstrojaner oder durch die Gewinnung von Hinter-grund-informationen (bspw. als Basis für die Durchführung von CEO-Fraud-Angriffen) erheblich größeren Schaden anzurichten.

Auswirkungen
Phishing- bzw. Social-Engineering-Angriffe durch Deepfake-Technologien sind besonders gefährlich, weil sie die Glaubwürdigkeit und Authentizität von **Kommunikationspartnern oder Nachrichteninhalten** mit relativ geringem Aufwand und in ziemlich kurzer Zeit untergraben können.

Der Massenversand von betrügerischen Nachrichten an potentielle Opfer wird besonders dann erfolgreich sein, wenn die eingehenden Nachrichten möglichst authentisch sind – und genau dabei helfen Deepfake-Technologien, denn sie erstellen manipulierte Inhalte, die schon heute kaum als solche identifiziert werden können.

Wenn zudem durch die gedeepfakten Inhalte Ängste von Menschen ausgenutzt oder Panik geschürt wird (wie bspw. durch die gefälschten Fotos zum Bombenangriff auf das Pentagon), dann haben diese Angriffe eine noch höhere Erfolgschance.

Bei auf einzelne Individuen abzielenden Angriffen können Deepfakes auf einer ganz anderen Ebene eingesetzt werden: Hier ist es wichtig, das Vertrauen des Opfers zu gewinnen. Auch das gelingt auf Basis realistisch gefälschter Medieninhalte wie Audio- oder Videostreams besonders gut und glaubhaft.

Da Deepfakes immer realistischer und schwerer zu erkennen sein werden, steigt das Risiko, dass Menschen diesen Angriffen zum Opfer fallen, derzeit erheblich.

Deepfake-Angriffstyp 6: Identitätsdiebstahl
Bislang spielte Identitätsdiebstahl als Angriffsszenario im unternehmensbezogenen Kontext nicht wirklich eine relevante Rolle. Denn hierbei handelte es sich unserer Erfahrung nach primär (natürlich nicht ausschließlich) bislang um ein Angriffsszenario, das im privaten Bereich und nicht so sehr im B2B-Umfeld von Relevanz war.

Durch das Aufkommen von Deepfakes hat sich diese Situation nun aber grundlegend geändert:

Denn auf Basis gedeepfakter Medieninhalte (v. a. Fotos, Audio- und Videosequenzen) besteht die sehr reale Gefahr, dass biometrische Identitätsverifizierungsverfahren (bspw. Stimmen- oder Gesichtserkennung) durch realistische Fälschungen erfolgreich getäuscht werden können und sich Angreifer so Zugriff auf bislang als sicher geltende Informationen und Ressourcen verschaffen können.

Durch den Einsatz generativer KI wird die Reproduktion biometrischer Eigenschaften im digitalen Bereich nun plötzlich mit sehr überschaubarem Aufwand möglich – und das macht biometrische Verfahren ungewohnt anfällig für Missbrauch. Diese neue Entwicklung wird schwerwiegende Folgen für die Sicherheit und Privatsphäre von Personen und Organisationen haben, die auf diese Verfahren vertrauen.

Bei auf Deepfakes beruhendem Identitätsdiebstahl geht es also um die Verwendung von durch generative Künstliche Intelligenz erzeugte biometrische Informationen, um die Identitätsverifizierungssysteme zu täuschen und sich als jemand anderes anzumelden.

Auswirkungen
Die Auswirkungen von Identitätsdiebstahl durch die Überlistung biometrischer Sicherheitssysteme im Unternehmensumfeld variieren erheblich von Unternehmen zu Unternehmen: So gibt es schon heute hochprofessionelle biometrische Sicherheitssysteme (dazu zählt bspw. auch die Gesichtserkennung bei Apple), die mindestens derzeit nicht durch Deepfakes überlistet werden können. Auf der anderen Seite sind aber durchaus ältere bzw. einfachere biometrische Sicherheitsverfahren im Einsatz, die durchaus überlistet werden können. Hier liegt es an den betroffenen Unternehmen, die eigene Situation

Identitätsdiebstahl

zu analysieren und, wo nötig, für Abhilfe zu sorgen, um das Schadensausmaß zu minimieren.

Deepfake-Angriffstyp 7: Betrug

Der Vollständigkeit halber (da im Unternehmenskontext maximal in Ansätzen relevant) sei noch ein weiteres Angriffsszenario erwähnt: allgemeine Betrugsmaschen auf Basis von Deepfake-Technologie. Hier steht nicht etwa – wie beim CEO-Fraud – die Führungsebene der Unternehmen im Fokus, sondern wir alle als Konsumenten bzw. Privatpersonen.

Wie bereits im Abschnitt 'Deepfakes – Einführung' an einigen Beispielen beschrieben, werden Deepfakes prominenter Personen wie Fernsehmoderator*innen und Politiker*innen auch dazu genutzt, für illegale Betrügereien wie falsche Investmentplattformen, etc. zu werben. Hier nutzen Kriminelle die Seriosität der gedeepfakten Personen aus, um Vertrauen bei den Empfängern dieser 'Werbesequenzen' zu generieren und diese damit dazu zu bewegen, Geld in ihre Vorhaben zu 'investieren'.

Hiervon betroffen sind aber nicht nur prominente Personen, sondern alle Personen, denen das Opfer vertraut – also bspw. auch Familienangehörige, Verwandte, etc.; so fällt die moderne Variante des Enkeltricks, in der gedeepfakte Stimmen von Familienangehörigen genutzt werden, um die Opfer dazu zu bewegen, in kürzester Zeit Geld an die Betrüger zu überweisen, bspw. ebenfalls in diese Kategorie.

Gefälschte Werbung mit vermeintlichen Aussagen von Prominenten oder von Vertrauten ist nicht neu – auch diese Form des Betrugs gibt es schon seit vielen Jahren. Aber auch in diesem Feld sorgt die Kombination der neuen technischen Möglichkeiten (bspw. zur einfachen, schnellen und kostengünstigen Generierung von täuschend echten bewegten Bildern und der Stimme des Prominenten/Vertrauten) mit der schnellen Verbreitung über die sozialen Medien dafür, dass diese Betrugsmasche völlig neue Dimensionen erreicht.

Auswirkungen

Wie eingangs bereits geschrieben, zielen diese Betrugsmaschen primär auf Konsumenten und nicht so sehr auf Unternehmen ab, insofern ist davon auszugehen, dass der hierdurch entstehende Schaden im Unternehmenskontext vglsw. gering sein wird.

SCHUTZMASSNAHMEN

Nüchtern betrachtet stellen Deepfake-Angriffe nur eine weitere Form von Cybersicherheitsrisiken dar; etwas realistischer betrachtet stellen sie eine erhebliche Steigerung des Sicherheitsrisikos für Unternehmen sowie auch für die Öffentliche Hand und Privatpersonen dar. Und zudem stellen

Schulungen

sie ein Risiko dar, dessen Schadensausmaße, wie oben näher ausgeführt, derzeit kaum realistisch abgeschätzt werden können.

Dennoch gilt auch und gerade unter Einbezug von Deepfake-Angriffsszenarien, dass die bislang eingeleiteten Schutzmechanismen noch immer sinnvoll sind und gelten, sie aber an bestimmten Stellen weiter ausgebaut und durch zusätzliche Inhalte/Maßnahmen ergänzt werden müssen.

Sensibilisierung

Sowohl gesamtwirtschaftlich als auch bezogen auf einzelne Unternehmen stellt die Sensibilisierung der Bevölkerung und der Mitarbeiter*innen einen, wenn nicht den absolut zentralen Ansatzpunkt dar:

Die präventive Aufklärung und Sensibilisierung möglichst vieler Personen zum Inhalt und Ablauf dieser neuen Angriffsszenarien ist extrem wichtig, damit sie in solchen Situationen

- sich nicht einschüchtern lassen (Erpressungsversuche)
- kritisch reagieren (CEO-Fraud, Phishing/Social-Engineering)
- von sich aus aktiv werden (bspw. von sich aus den vermeintlichen 'Auftraggeber' – also CEO, Geschäftsleitung, etc. – auf einem anderen Kanal aktiv kontaktieren) (CEO-Fraud)
- sich proaktiv auf das Eintreffen eines Angriffs vorbereiten und sinnvolle Gegenmaßnahmen planen können

Darüber hinaus ist es wichtig zu vermitteln, wie leichtgewichtig und schnell Deepfake-basierte Manipulationen auch für technische Laien – und erst recht für Kriminelle – herzustellen sind und wie realistisch diese Fälschungen mittlerweile sind.

Sofern Unternehmen, die mit derartigen Sensibilisierungs- und Schulungsmaßnahmen verbundenen Kosten und zeitlichen Aufwände nicht schultern können bzw. wollen, sollten sie mindestens diejenigen Mitarbeiter*innen aufklären, die für sensible Bereiche wie Buchhaltung, Finanzen oder Personal zuständig sind.

Die Ausführungen im Abschnitt zu Erpressungsangriffen haben aber hoffentlich ein-

drücklich unterstrichen, welche Gefahren von allen Mitarbeiter*innen ausgehen – insofern ist es (fast) fahrlässig, hier aufgrund von Kosten- oder Zeitargumenten nicht möglichst zeitnah das gesamte Personal entsprechend aufzuklären und diese Maßnahmen regelmäßig zu wiederholen und zu vertiefen.

Schulung

Neben der grundlegenden Sensibilisierung für solche Angriffsszenarien kann den Teilnehmenden in darüberhinausgehenden Schulungen/Workshops auch vermittelt werden, wie sie Anzeichen von Angriffen oder/und Manipulation erkennen können, z. B. ungewöhnliche Anfragen, Dringlichkeit, Artefakte, Sprachfehler oder Hintergrundgeräusche.

Solche Schulungen sind zwar durchaus eine sinnvolle Ergänzung, aber im Gegensatz zu den oben beschriebenen Sensibilisierungsmaßnahmen keine notwendige Schutzmaßnahme.

Notfallplan und Krisenmanagemen

Wie in allen iT-sicherheitsbezogenen Kontexten gilt auch bei Deepfake-Angriffen: Ein extrem wichtiger reaktiver Schutz ist ein professioneller und ausgearbeiteter Notfallplan und ein entsprechendes Krisenmanagement.

Wenn im Schadenfall jede*r genau weiß, was sie*er zu tun hat, wer Ansprechpartner*innen sind und wie mit der Presse, der Öffentlichkeit, Partnern und Kunden umgegangen werden sollte, ist das ein Garant dafür, dass der entstandene Schaden minimiert werden kann.

Klare Verfahrensregeln

Gerade gegen Angriffe wie (Deepfake) CEO-Fraud oder Betrug ist eine der sinnvollen und wichtigsten Schutzmaßnahmen natürlich die Einführung und Einhaltung klarer Verfahren für Finanztransaktionen, die eine konsequente Mehrfachauthentifizierung und -genehmigung erfordern.

Natürlich kann auch eine Mehrfachgenehmigung durch Deepfake-Angriffe ausgehebelt werden, indem eben nicht nur eine Person, sondern zwei getäuscht oder imitiert werden. Dennoch senken klare Verfahrensregeln mit Mehrfachauthentifizierung und Mehrfachgenehmigung das Risiko, Opfer solcher Angriffe zu werden, ganz erheblich.

Technische Schutzmaßnahmen

Neben den organisatorischen Maßnahmen sind – wie immer – selbstverständlich auch technische Maßnahmen sinnvoll und wichtig, um das durch Deepfake-Angriffe entstehende Risiko effektiv zu minimieren.

Natürlich wäre die möglichst schnelle Einführung von Werkzeugen, die dabei helfen, manipulierte Inhalte zu erkennen bzw. die Authentizität echter Medieninhalte zu bescheinigen, eine sehr effektive Schutzmaßnahme.

Stand heute gibt es allerdings (noch) keine hinreichend-ausgereiften Tools, die bspw. durch den Einsatz von Künstlicher Intelligenz, Blockchain oder digitalen Wasserzeichen dabei helfen könnten, die Echtheit übermittelter Fotos, Audio- oder Videoinhalte zu verifizieren. Solche Lösungen wären ideal und sind in der Entwicklung, aber es ist noch nicht absehbar, wann sie zum Standard in der digitalen Kommunikation werden; zudem lassen sich nicht alle Formen von Deepfake-Angriffen durch sie abwehren.

Die Nutzung vertrauenswürdiger Kommunikationskanäle und -plattformen, die eine sichere Verschlüsselung und Identitätsüberprüfung bieten, um die Wahrscheinlichkeit von gefälschten Anrufen oder Nachrichten zu reduzieren, stellt einen weiteren wichtigen Ansatz dar.

Gerade im Umfeld von biometrischen Verfahren (Identitätsdiebstahl) sollten dort, wo dies ggf. noch nicht geschehen ist, robustere und zuverlässigere Methoden der Identitätsverifizierung entwickelt werden. Dies kann u. a. dadurch geschehen, dass Features wie Verfahren zur Lebendigkeitserkennung einbezogen und weiterentwickelt werden; andererseits sollten im Sinne einer Multi-Faktor-Authentifizierung Ansätze implementiert werden, die auf einer Kombination aus biometrischen und nicht-biometrischen Verfahren wie Wissen oder Besitz basieren.

Juristische Rahmenbedingungen

Nur der Vollständigkeit halber sei erwähnt, dass natürlich auch der Gesetzgeber einen entsprechenden Rahmen auf europäischer und internationaler Ebene schaffen muss, in dem klar geregelt wird, dass Deepfake-basierte Angriffe rechtswidrig sind und mit entsprechenden Strafen geahndet werden.

ERGEBNISSE UND AUSBLICK

'Deepfakes' sind ein vglsw. junges Phänomen (lt. Wikipedia wurde der Begriff erstmals 2017 genutzt). Wie in den vorangegangenen Abschnitten deutlich geworden sein dürfte, hat sich diese Technologie dennoch allein in den letzten 1,5 Jahren durch die enormen Fortschritte im Bereich der generativen Künstlichen Intelligenz und der Leistungssteigerung moderner Hard- und Software massiv weiterentwickelt.

Während es viele nutzbringende Anwendungen für Deepfakes gibt, führt ihr Einsatz im Bereich der Cyberkriminalität zu Risiken, deren Auswirkungen wir heute u. E. noch gar nicht ansatzweise abschätzen können. In diesem Zusammenhang ist es wichtig zu erkennen, dass derzeit die potentielle Erpressbarkeit aller Mitarbeiter*innen eine der größten Bedrohungen darstellt. Hier reichen oft problemlos herzustellende gedeepfakte Fotos oder Stimmaufnahmen aus, um eine hinreichende Erpressungsgrundlage zu erhalten.

Schon heute befinden wir uns in einer Situation, in der Deepfakes so weit fortgeschritten sind, dass Menschen und sogar Systeme an vielen Stellen (mindestens im Alltagstrubel) nicht mehr in der Lage sind, zwischen Realität und Illusion zu unterscheiden. Aufgrund des exponentiellen technischen Fortschritts gilt als sicher, dass sich dieser Trend in den nächsten Monaten fortsetzen wird und es immer schwieriger bis unmöglich wird, manipulierte Inhalte von der Realität zu unterscheiden. Dieser besorgniserregenden Entwicklung werden wir erst Einhalt bieten können, wenn neue technische Werkzeuge geschaffen und standardmäßig auf einer breiten Basis eingesetzt werden, um die Authentizität von Medieninhalten eindeutig nachzuweisen.

Zugegeben: Heute (also Mitte 2024) sind viele Deepfakes noch von minderer Qualität und daher vglsw. leicht erkennbar; werden noch viele Trainingsdaten für die Erstellung gedeepfakter Inhalte benötigt; ist es noch schwer, an wirklich gute Deepfake-Tools heranzukommen; ist die Erstellung qualitativ-hochwertiger, manipulierter Medieninhalte (v. a. Videoaufnahmen) noch mit erheblichen Kosten verbunden.

Aber: Schon heute existieren Werkzeuge und Forschungsprojekte, welche erheblich weniger Audio- und/oder Videomaterial benötigten, um absolut realistische Deepfakes zu erstellen; zudem entwickelt sich der technische Fortschritt – wie oben mehrfach beschrieben – in rasantem Tempo weiter und bringt uns damit immer neue Werkzeuge, die gerade Kriminelle im Umfeld von Deepfakes einsetzen können; und wir sollten nicht vergessen, dass wir im beruflichen Alltag nicht immer die Zeit und Möglichkeiten haben, in Ruhe – geschweige denn als Nicht-iT-Profi das Wissen und die Werkzeuge haben – Medieninhalte zu analysieren; und schließlich auch, dass es genügend Rahmenbedingungen wie Mobilfunk, kleine Handybildschirme, geringe Übertragungsgeschwindigkeiten, etc. gibt, bei denen auch echte Medieninhalte in schlechter Qualität übertragen werden – was eine Beurteilung selbst vermeintlich schlechter Deepfakes auch heute schon erheblich erschwert.

Aber es sind nicht nur professionelle Kriminelle, die diese Tools nutzen werden – da sich die benötigte Expertise und der notwendige Aufwand zur Erstellung von Deepfakes durch neue Tools infolge des technischen Fortschritts stetig verringern werden, werden auch technische Laien (bspw. Kinder/Teenager) problemlos Deepfakes produzieren können. Damit ist schon heute absehbar, dass die Häufigkeit und damit verbunden auch die Auswirkungen von Deepfake-basierten Angriffen in der nahen Zukunft massiv steigen werden.

Vielleicht viel wichtiger: Das Wissen um diese neue Technologie und die neuen Möglichkeiten sowie Risiken verbreitet sich noch erheblich zu langsam in allen Teilen der Bevölkerung und auch in der Wirtschaft – dies macht uns alle unnötig anfälliger für diese Risiken, die ohnehin in vielen Fällen schon existenzbedrohend werden können.

Wir hoffen, dass wir mit dieser Publikation und unseren Workshops einen kleinen Teil dazu beitragen können, diesen Missstand zu beheben und gemeinsam dafür zu sorgen, dass wir mehr Nutzen als Schaden aus dieser beeindruckenden neuen Technologie ziehen werden.

Autoren:
TOBIAS RADEMANN
Geschäftsführer
NIKLAS ZISTLER
Head of iT-Security

R.iT GmbH
Lise-Meitner-Allee 37
44801 Bochum
Telefon: +49 (234) 43 88 00-0
Telefax: +49 (234) 438800-29
eMail: info@RiT.de
Internet: www.RiT.de

Fotos: iStockphoto (Titel), Adobe Stock