Alles Lüge

Künstliche-Intelligenz-Werkzeuge für Bilder und Videos sind mittlerweile frei verfügbar, so gut wie jeder kann damit Fälschungen erstellen – auch für Zwecke wie Betrug, Propaganda und Mobbing. Die Flut dieser sogenannten Deep Fakes könnte die Wahrnehmung von Medien im Internet grundsätzlich verändern

Dieses Foto des früheren US-Präsidenten Donald Trump mit schwarzen Frauen und Männern ist gefälscht und wurde in sozialen Netzwerken verbreitet

Anfang des Jahres kursierte auf der Social-Video-Plattform TikTok ein kurzer Clip von Ricarda Lang im Gespräch mit Talkshow-Moderator Markus Lanz, in dem die Grünen-Vorsitzende auf die Frage „Was wäre denn, wenn die Bauern uns jetzt keine Kartoffeln mehr liefern?“ antwortet: „Das wäre mir, ehrlich gesagt, egal, da ich eh nur Pommes esse.“ Das Video erschien nur auf den ersten Blick echt, sammelte bei TikTok und Instagram dennoch viele Likes ein. Natürlich hatte Lang die Lanz-Frage in Wirklichkeit so nie beantwortet. Stattdessen hatte ein unbekannter TikTok-Nutzer die Talkshow-Aufzeichnung mit einem Werkzeug zur Manipulation von Bewegtbildern per künstlicher Intelligenz (KI) bearbeitet.

VON BENEDIKT FUEST

Das Lang-Video ist nur eines von vielen sogenannten „Deep Fakes“, die derzeit in sozialen Netzwerken kursieren und auf halbseidenen Netzplattformen verbreitet werden: Der britische Premier Rishi Sunak wirbt etwa auf Facebook per Video angeblich für eine Gruppe von Anlagebetrügern, falsche Nacktfotos spanischer Schülerinnen kursieren in WhatsApp-Gruppen, und auf Telegram werden vermeintliche Porno-Aufnahmen der US-Stars Taylor Swift und Selena Gomez verbreitet. Jedes der Videos und Fotos wurde per KI gefälscht, nichts von dem Gezeigten ist jemals so real passiert. Die Werkzeuge dafür können die Täter inzwischen nicht nur im Internet herunterladen, sondern auch sehr einfach anwenden. Ihre freie Verfügbarkeit werde die Art, wie Medien wirken, komplett verändern, warnen Experten.

Auf der Münchner Sicherheitskonferenz Mitte Februar war die Beeinflussung der kommenden US-Wahlen mittels gefälschter Videos ein bestimmendes Thema. Manager der Technologiekonzerne Adobe, Amazon, Google, IBM, Meta, Microsoft, OpenAI und TikTok einigten sich auf gemeinsame Maßnahmen zur Kennzeichnung und Erkennung von KI-generierten Bildern und Videos auf ihren Plattformen. Doch der Ansatz bleibt bislang unkonkret und fast hilflos. Denn eine Kennzeichnung von KI-Bildmaterial etwa per Wasserzeichen funktioniert nur, wenn alle Bildgeneratoren die Funktion auch einsetzen. Das aber ist längst nicht mehr der Fall.

MASSENWEISE VERBREITET

„Die Betreiber der großen KI-Bildgeneratoren wie Midjourney oder OpenAIs Sora haben zwar schon früh Sperren in ihre Algorithmen eingebaut“, sagt Rüdiger Trost vom finnischen IT-Sicherheitsunternehmen WithSecure. „Doch Open-Source-KI-Werkzeuge werden immer weiter verbessert und weit verbreitet eingesetzt.“ Open Source bedeutet, dass eine Software für jedermann zugänglich ist. Inzwischen könnten auch Laien frei verfügbare KI-Modelle einfach herunterladen und trainieren, erklärt der Experte. „Dazu benötigt man nur Trainingsmaterial und einige Stunden Rechenleistung eines leistungsfähigen Computers – die kann man zum Beispiel einfach in der Amazon-Cloud mieten.“ Wer erst einmal einen Algorithmus trainiert hat, sei es mit Fernsehauftritten von Sunak, Bildern von Swift oder Videos einer Porno-Webseite, kann anschließend selbst passende Deep Fakes erstellen. „Selbst das Material aus zehn Minuten unseres Gespräches würde bereits ausreichen, um Ihre Stimme zu fälschen“, warnt Trost.

Diese Möglichkeiten nutzen Kriminelle bereits häufig. In Hongkong etwa überwies der Buchhalter eines britischen Finanzunternehmens umgerechnet 22 Millionen Euro auf die Konten von Betrügern, nachdem er in einer Videokonferenz mit mehreren Kollegen vom Finanzchef der Firma dazu angewiesen wurde. Laut dem Polizeibericht stammten jedoch sämtliche Videobilder der Vorgesetzten und Kollegen von einem Echtzeit-Algorithmus. Auch Sunaks unfreiwillige Werbeauftritte kamen laut einer Analyse der Desinformationsexperten der britischen Beratung Fenimore Harper von einem Open-Source-KI-Werkzeug. Sie warnten im Februar bei einer Anhörung vor dem Parlament in London vor KI-Inhalten, die ohne jede menschliche Aufsicht massenweise auf sozialen Netzwerken verbreitet werden und dort die Meinung prägen.

Angesichts dieser Fälle ist fraglich, inwieweit KI-Gesetze die Flut der Deep Fakes noch eindämmen können. Europa versucht es dennoch: In der vergangenen Woche stimmte das europäische Parlament für das KI-Gesetz der Union. In dessen Artikel 52, Absatz 3 werden KI-Anwender dazu verpflichtet, „Deep Fakes (...) als künstlich erzeugt oder manipuliert“ zu kennzeichnen. Eine Haftung für Entwickler der entsprechenden Werkzeuge oder gar für Social-Media-Plattformen, auf denen „Deep Fakes“ verbreitet werden, sieht das Gesetz jedoch nicht vor.

Ohnehin ist zweifelhaft, inwieweit die bloße Kennzeichnung dazu geeignet ist, die Verbreitung und Popularität der Fälschungen zu bremsen. Bereits jetzt verdienen mehrere große Deep-Fake-Pornoplattformen mit unechten Filmen auf der Basis von Prominenten-Bildmaterial ihr Geld. Sie sind über Anonymisierungsdienste unter Internetadressen von kleinen Inselstaaten im Pazifik erreichbar, wo die westliche Justiz wenig Handhabe hat.

Deswegen versuchen die Opfer von Deep Fakes in den USA aktuell per Copyright-Beschwerden, ihre Bilder zumindest aus den Suchergebnissen im Internet zu entfernen. Betroffene reichten bislang mehr als 13.000 Beschwerden wegen Deep-Fake-Pornos bei Google ein. Parallel versuchen Anwälte von Prominenten, den Internetriesen dazu zu bewegen, die Seiten ganz zu sperren.

ÜBER EINE SMARTPHONE-APP

Doch nicht nur Stars trifft es. Im Herbst 2023 musste eine Gruppe von 20 Schülerinnen im Alter zwischen elf und 17 Jahren aus dem spanischen Städtchen Almendralejo feststellen, dass ihre Mitschüler eine frei verfügbare Smartphone-App dazu genutzt hatten, sie virtuell auszuziehen. Die mittels künstlicher Intelligenz generierten Nacktfotos wurden monatelang über WhatsApp-Gruppen verbreitet. Dass sie offensichtlich falsch waren, tat ihrer Popularität keinen Abbruch. Erst als die Ärztin Miriam Al Adib, Mutter eines der Mädchen, per Instagram auf den Fall aufmerksam machte, wurden die spanischen Datenschutz-Behörden aktiv, die Bilder verschwanden aus den Netzwerken und von den Smartphones.

„Diese Art von Deep-Fake-Bildern von Jugendlichen sind eine neue digitale Spielart des alten Mobbing-Problems“, sagt der Kommunikationswissenschaftler Alexander Godulla von der Universität Leipzig. Er forscht im Rahmen des „Deep Fake“-Projekts an der Universität zum Phänomen KI-Fälschungen und hält eine Kennzeichnungspflicht für wenig hilfreich. Dass die offensichtlich unechten freizügigen Bilder von Taylor Swift ausgerechnet zu dem Zeitpunkt erschienen, als der Hype um eine mögliche politische Dimension ihrer Beziehung mit einem US-Football-Star neue Höhen erreichte, hält Godulla für keinen Zufall. „Es geht den Tätern darum, virtuell in die Intimsphäre von Frauen einzudringen, sie zu erniedrigen und aus dem öffentlichen Diskurs zu verdrängen“, erklärt er. „Nicht umsonst gibt es bereits viele entsprechende Deep Fakes von Politikerinnen – tatsächlich erotisches Material zu generieren ist da nicht das Ziel der Täter.“

Dass das Material dennoch weiterverbreitet wird, beweise einen Verfall der Debattenkultur in sozialen Medien, schreibt die US-Medienforscherin Renée DiResta vom „Internet Observatory“ der kalifornischen Stanford-Universität. „Dieser soziale Niedergang überschneidet sich mit zwei anderen Trends“, meint sie. „Soziale Medien senken die Kosten der Verbreitung gefälschter Inhalte, generative KI reduziert die Erstellungskosten. Der Belästigungsinhalt ist interessanter und eindringlicher als irgendein überflüssiges Meme.“ Wirkungsvoller also als ein kleines lustiges Bildchen oder Kurzvideo.

TECHNISCHE LÖSUNG

Der Leipziger Forscher Godulla sagt, Nutzer seien eher bereit, offensichtlich gefälschte Inhalte weiterzuverbreiten, wenn sie ihrem politischen Weltbild entsprächen. Darin aber liege die größte Gefahr für aktuelle Wahlkämpfe, etwa in den USA. „Nicht nur die Qualität der Fälschungen ist neu, sondern vor allem die Quantität“, erklärt er. „Wir haben plötzlich ganz viele Deep Fakes, und jedes dieser Deep Fakes verlangt von den Nutzern eine Reflektion über die Botschaft: Kann das wahr sein? In der Breite sind die Menschen dafür nicht kompetent genug.“

Stattdessen könnte die Flut der Deep Fakes die Glaubwürdigkeit von Medien insgesamt unterminieren, fürchtet der Wissenschaftler. „Wenn alles falsch sein kann, glaubt man am Ende gar nichts mehr.“ Eine technische Lösung könnte darin liegen, Bilder bereits in der Kamera mit Wasserzeichen zu versehen, die in Redaktionen klassischer Medien und Nachrichtenagenturen verifiziert werden. Professionelle Medien könnten dank dieser Kennzeichnung gegenüber ihren Lesern die Authentizität der Bilder und Videos garantieren. Entsprechende Kamera-Technik gibt es bereits. „Für ein medienkompetentes Publikum ist das eine Chance“, so Godulla. „Doch wer sich ohnehin nur noch über soziale Medien mit Nachrichten versorgt, dem ist ein Wasserzeichen auch gleichgültig.“ Deep Fakes, so seine Sorge, werden in Zukunft politische Grabenkämpfe nur verschärfen, da sich jeder lediglich die Bilder aussucht, die seine Meinung bestätigen.