Wie Gesichter gestohlen werden

Das FBI warnt vor Deepfakes, Facebook behauptet, eine Lösung zu haben. Entstanden sind die computergenerierten Fälschungen als Kollateralschaden einer klugen Idee. VON RUTH FULTERER

Künstlich erzeugte Videos, Bilder und Tonaufnahmen werden zu einem Sicherheitsproblem. Und das wohl schon in den nächsten Monaten. Das ist zumindest die Einschätzung des FBI: In Zukunft würden fremde Geheimdienste und Kriminelle immer öfter computergenerierte Fälschungen nutzen, um in Computernetzwerke einzudringen, schreibt es in einer Meldung.

Gemeint sind sogenannte Deepfakes, also von Computern gemachte, täuschend echte Fälschungen. Das können Texte und Bilder in Phishing-E-Mails sein, aber auch gefälschte Sprachnachrichten – scheinbar vom Chef, der einen Notfall habe und dringend Geld brauche. Solche Fälschungen sind nicht nur ein Einfallstor für Hacker. Sie machen auch Fake-News-Kampagnen glaubhafter oder helfen Betrügern etwa dabei, mit gefälschten Bildern von Schäden Versicherungen hinters Licht zu führen.

Noch wird die sogenannte Deepfake-Technologie vor allem dazu eingesetzt, Köpfe von prominenten Frauen auf jene von Pornodarstellerinnen zu setzen. 2019 waren laut einem Report der Cyber-Security-Firma Deeptrace 96 Prozent aller Deepfake-Videos im Internet pornografisch.

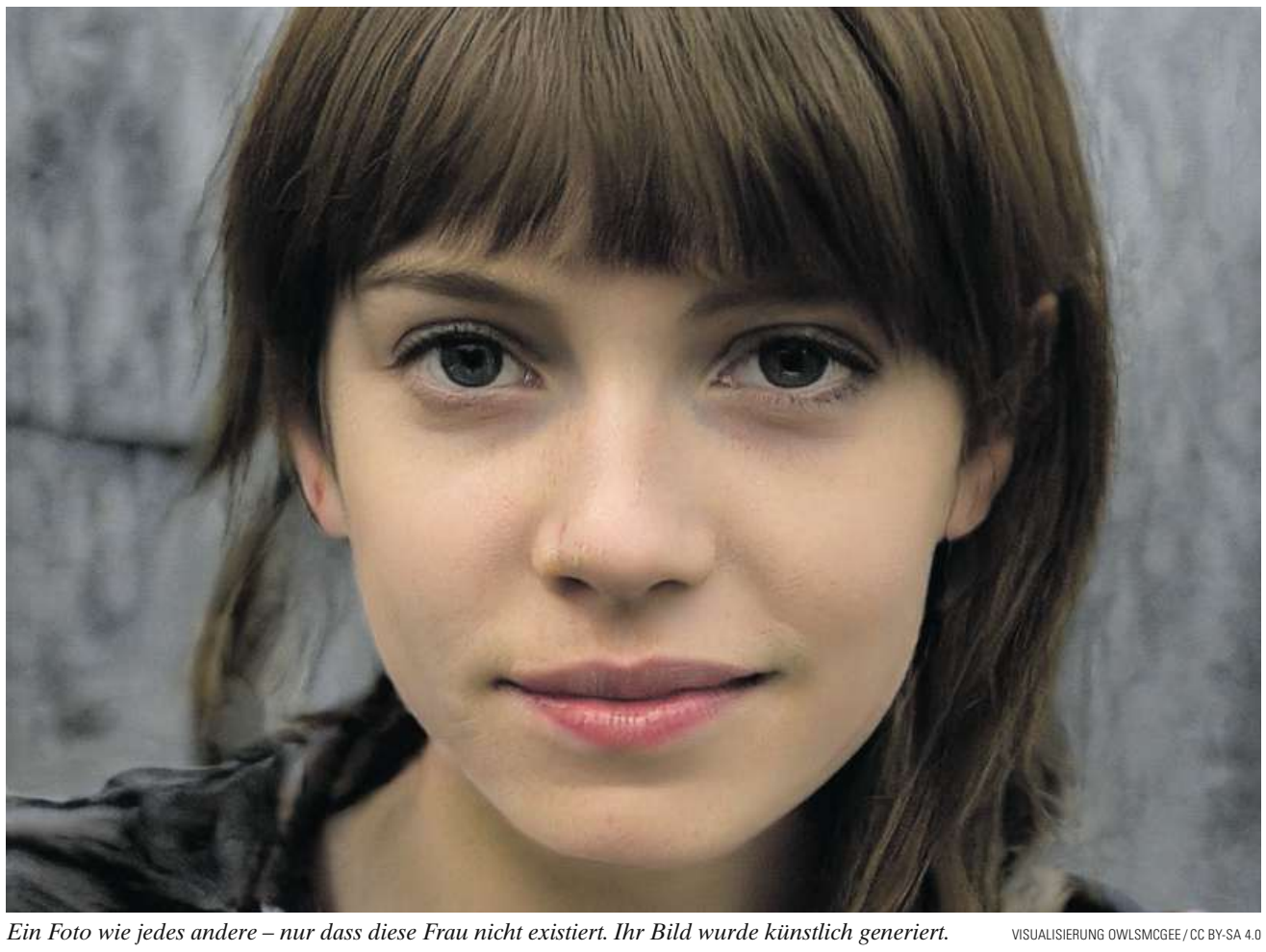
Teilweise sollen diese Videos Frauen gezielt schaden, wie die Geschichte von Rana Ayyub zeigt: Die indische Journalistin hatte angeprangert, dass die nationalistische Partei BJP einen Kinderschänder verteidigte. Ihre Feinde starteten eine Hasskampagne gegen sie, unter anderem mit einem gefälschten Video, das sie beim Geschlechtsverkehr zu zeigen schien. Es wurde viral geteilt, und der Journalistin schlug so viel Wut und Empörung entgegen, dass sie für Monate nicht mehr öffentlich auftreten und ihre Arbeit machen konnte.

Das war im Frühling 2018. Seither ist die Fälschungssoftware immer besser geworden: Die Bilder sind realistischer, immer weniger Bildmaterial der «Zielperson» ist nötig. Nicht nur Prominente sind mögliche Opfer, sondern auch Privatpersonen.

Das alles hatte sich Ian Goodfellow wohl nicht träumen lassen, als er auf jene Idee kam, welche diese Entwicklungen überhaupt erst möglich machte. Das war im Jahr 2014. Goodfellow war damals Doktorand der University of Montreal. Bei einer Abschiedsparty eines Kollegen diskutierte er mit Freunden über ein grundlegendes Problem von selbstlernenden Algorithmen. Diese waren damals schon sehr gut darin, etwa Gesichter oder Tiere auf Bildern zu erkennen, allerdings erst, nachdem sie aufwendig trainiert worden waren, mit Millionen von beschrifteten Bildern. Solche Bilder samt Beschriftung kann man aus dem Internet absaugen, doch die Qualität ist oft dürftig. Goodfellows Freunde arbeiteten an einer Lösung. Er fand ihre Idee schlecht. Dabei sei er auf eine eigene gekommen, erzählte Goodfellow gegenüber Medien. Wenig später veröffentlichte er sie als Paper. Er nannte die Sache Generative Adversarial Networks (GAN).

Algorithmen im Wettbewerb

Die Grundidee von GAN ist, zwei Algorithmen gegeneinander antreten zu lassen. Der erste hat zum Ziel, zu erkennen, was ein falsches Bild ist und was ein echtes. Der zweite generiert möglichst realitätsnahe Bilder und wird belohnt, wenn er es schafft, den ersten zu täuschen. Dem Detektivsystem werden echte Bilder und solche vom Täuscher vorgelegt, und es wird belohnt, wenn es Bilder des Täuschers aufdeckt. Eingebaute Rückkopplungsschleifen führen dazu, dass sich die Systeme aneinander abarbeiten und so verbessern: Je besser der Täuscher, desto besser der Detektiv und umgekehrt. Das Ergebnis sind Bilder, die ein Computersystem generiert hat, die aber so echt aussehen, dass ein Mensch den Unterschied kaum oder gar nicht mehr erkennt.



Ein Foto wie jedes andere – nur dass diese Frau nicht existiert. Ihr Bild wurde künstlich generiert. VISUALISIERUNG OWLSMCGEE / CC BY-SA 4.0

Auch dieses System braucht beschriftete Daten, um ins Rollen zu kommen, doch sehr viel weniger. Für Goodfellow und viele andere Forscher in dem Bereich war das praktisch, denn sie konnten einen anfangs kleineren Datensatz künstlich mit täuschend ähnlichen Inhalten vermehren, ob es nun Bilder, Tonaufnahmen oder Videosequenzen waren. Goodfellow wurde bekannt. Inzwischen leitet er beim Tech-Konzern Apple die Abteilung für maschinelles Lernen. Die andere Folge der Technologie ist weniger erfreulich: Deepfakes.

Gesichter im Fake-Kontext

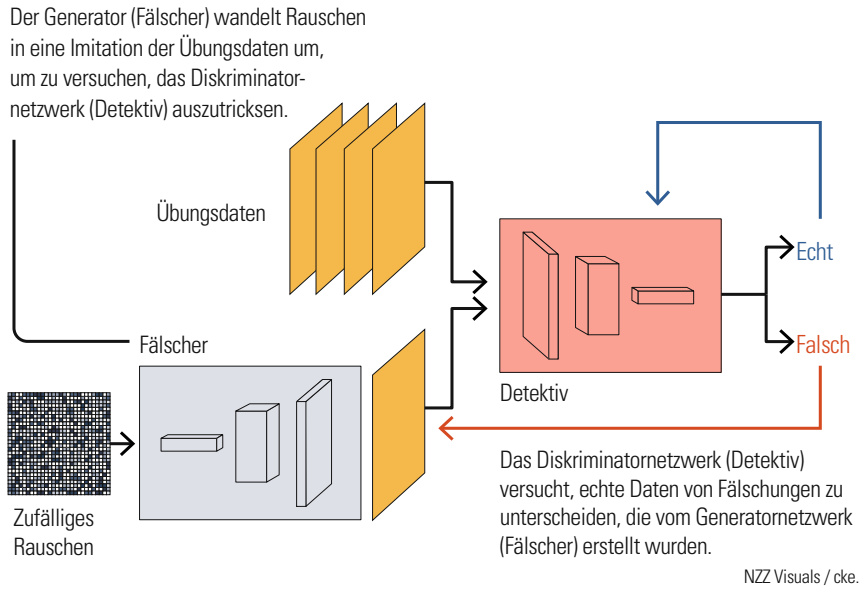
Vielleicht noch beunruhigender als Bilder von Menschen, die nicht existieren, sind jene, die reale Personen in falsche Kontexte bringen. Wer schon einmal mit einer App die Gesichter zweier Menschen ausgetauscht hat oder ein Gesicht künstlich altern liess, der weiss, dass schon ein einziges Originalbild zu realistischen Resultaten führen kann.

Für diese Art von Fakes ist noch ein weiterer Baustein nötig. Und zwar die sogenannten Autoencoder. Sie wurden erdacht, um eine Menge von Informationen auf das Notwendigste zu reduzieren. Aus einem Absatz können solche Modelle beispielsweise die wichtigsten Sätze und Stichworte herausziehen, aus dem Bild von einem Menschen die Gesichtszüge. Diesen Prozess nennt man Encoding. Man kennt es auch vom Umwandeln von Musikaufnahmen in Formate, die weniger Platz brauchen, etwa MP3.

Das faszinierende bei den Autoencoding-Algorithmen ist, dass der passende, mittrainierte Decoder-Algorithmus aus der abstrahierten Information wieder etwas herstellen kann, was dem ursprünglichen Bild, Absatz oder Audiostück sehr nahe kommt. Wenn man nun zwei Gesichter austauschen will, dann reicht es, die Gesichtszüge mit demselben Programm zu extrahieren, dann aber die Information zur Wiederherstellung zu vertauschen. Autoencoder-Modelle gibt es seit den 1980er Jahren. Doch erst in Kombination mit GAN entfalteten sie ihr Potenzial als Fälschungsalgorithmen.

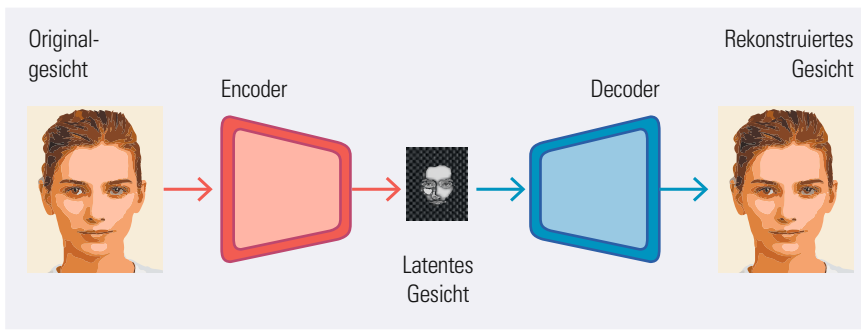
Nicht nur die Fortschritte der Technik machen Deepfakes zu einer immer grösseren Gefahr. Wir werden auch verwundbarer durch die zunehmende Digitalisierung. Was zum Beispiel, wenn der Auftritt eines Politikers bei einer virtuellen Konferenz in Echtzeit gefälscht würde?

Wie der Wettstreit zweier Algorithmensysteme Fälschungen hervorbringt



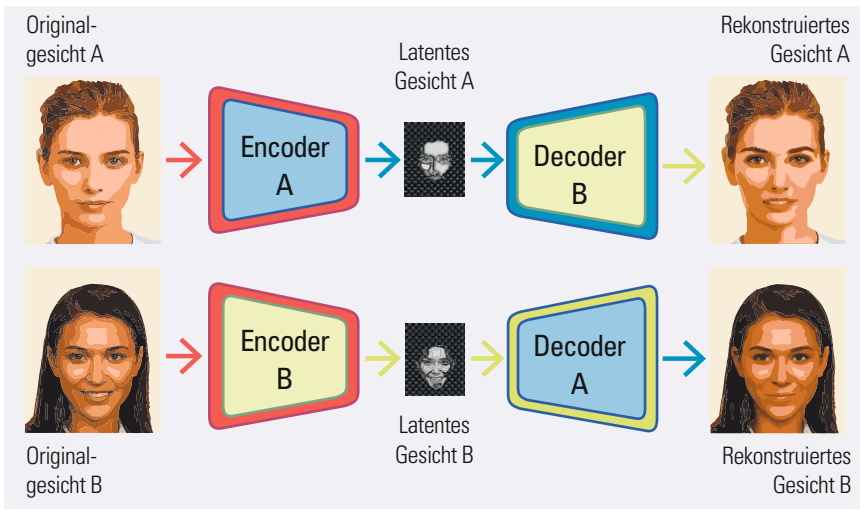
NZZ Visuals / cke.

Autoencoder-Modelle speichern zentrale Merkmale eines Bilds



QUELLE: T. T. NGUYEN ET AL. NZZ Visuals / cke.

Vertauschte Decoder ergeben Misch-Gesichter



QUELLE: T. T. NGUYEN ET AL. NZZ Visuals / cke.

> «Bei gut gemachten Fakes hat der Mensch keine Chance. Das sollte jedem klar sein, der gesehen hat, wie Luke Skywalker in ‹Star Wars› 30 Jahre jünger gemacht wurde.»
>
> **Anthony Sahakian**
> CEO von Quantum Integrity

Entwicklerinnen und Ingenieure arbeiten laufend an Software, die Fakes erkennen soll. Vor allem Facebook reagiert auf den Vorwurf, es helfe der Verbreitung von Falschnachrichten: Forscher des Konzerns und der Michigan State University haben einen Algorithmus darauf trainiert, in Fake-Bildern Hinweise auf die Erzeuger-KI zu finden, eine Art Fingerabdruck. Dieser würde es erlauben, alle Fakes aus einer Quelle auf einmal herauszufiltern. Die Ergebnisse seien vielversprechend. Die Latte liegt allerdings nicht sehr hoch. Bei einem Programmierwettbewerb zur Deepfake-Entlarvung erkannte 2019 selbst das beste System nur drei Viertel der gefälschten Videos.

Könnte das der Mensch besser? Das FBI rät in seiner Meldung, auf Pupillen oder Ohrläppchen zu achten, diese stelle Fälschungssoftware oft nicht wahrheitsgetreu nach. Doch der Experte Anthony Sahakian winkt ab: «Bei gut gemachten Fakes hat der Mensch keine Chance. Das sollte jedem klar sein, der gesehen hat, wie Luke Skywalker im letzten ‹Star Wars›-Film 30 Jahre jünger gemacht wurde.» In weniger als einem Jahr werde allgemein verfügbare Software eine ähnlich hohe Qualität erreichen, sagt er voraus. Sahakian führt die Genfer Firma Quantum Integrity, die Technologie zum Entdecken von digitalen Fälschungen entwickelt. Das ist ein Wettlauf gegen die Zeit. Die Software, die seine Firma im Moment verkauft, wird bald obsolet sein. Er sagt: «Fälscher- und Aufdecker-Software übertrumpfen sich ständig gegenseitig, wie Computerviren und Firewalls.»

Blockchain als Hoffnung

Wie kann man sich also schützen? Allgemein steigt mit jeder Ebene an Information die Hürde für Fälscher. Ein Bild allein fälscht sich leichter als ein Video. Videos mit Ton sind schwerer zu fälschen als solche ohne; vor allem Stereo-Sound lässt sich nur schwer maschinell erzeugen. Bei Bildern, die etwas beweisen sollen, sollte man dazu mehr auf Metadaten achten, etwa die GPS-Daten einer Aufnahme. Solche Zusatzinformationen werden in Zukunft wohl vermehrt auf verschlüsselten Datenbanken oder Blockchains gespeichert werden. Der Scan eines Personalausweises allein verliert an Wert, eingebaute Chips mit biometrischen Informationen und besondere Drucktechniken gewinnen an Bedeutung.

Gegen so manche Fälschung würde auch eine skeptische Grundhaltung helfen. Das Problem dabei: Je eher eine Information in das eigene Weltbild passt, desto eher glauben Menschen daran. Auch das hat die Fake-News-Epidemie gezeigt: Die Qualität der Fälschungen muss gar nicht überragend sein, solange sie Menschen in der eigenen Meinung bestärken.