CAPCO

DEEPFAKES

& DISINFORMATION

a wipro company

INHALTSVERZEICHNIS

EXECUTIVE SUMMARY

Der Einsatz Künstlicher Intelligenz (KI) spielt eine immer größere Rolle in unserer Gesellschaft – mit den neuen Möglichkeiten dieser Technologie gehen aber auch neue Risiken einher. Eines dieser Risiken ist der Missbrauch der Technologie für die bewusste Verbreitung von falschen Informationen. Politisch motivierte Desinformation ist gewiss kein neues Phänomen, doch der technologische Fortschritt macht die Erzeugung und Verbreitung manipulierter Inhalte deutlich einfacher und effizienter als früher. Mithilfe von KI-Algorithmen lassen sich heutzutage Videos schnell und relativ günstig fälschen (Deepfakes), ohne dass hierfür noch Spezialkenntnisse erforderlich wären.

Wenn auch in erster Linie über den möglichen Einsatz von Deepfakes in Wahlkampagnen diskutiert wird, so macht diese Art von Videos doch nur einen Bruchteil aller Manipulationen aus: In 96 Prozent der Fälle werden Deepfakes dazu genutzt, pornografische Filme mit prominenten Frauen zu erzeugen. Auch Frauen, die nicht in der Öffentlichkeit stehen, finden sich als unfreiwillige Hauptdarstellerinnen in manipulierten Videos wieder (Deepfake-Rachepornografie). Zudem können statische Bilder mit Anwendungen wie DeepNude in täuschend echte Nacktbilder umgewandelt werden. Wenig überraschend funktionieren diese Anwendungen nur mit Bildern von Frauenkörpern. Aber nicht nur visuelle Inhalte können manipuliert oder algorithmisch produziert werden. KI-generierte Stimmen wurden bereits erfolgreich für Betrug mit hohen finanziellen Schäden angewendet, und mit GPT-2 können Texte erzeugt werden, die Fakten und Zitate beliebig erfinden.

Wie sollten wir mit dieser Herausforderung am besten umgehen? Unternehmen und Forschungsinstitute investieren bereits hohe Summen in technologische Lösungen, um KI-generierte Videos zu identifizieren. Der Nutzen dieser Investitionen ist meist nicht von Dauer: Sobald die Ergebnisse publik werden, können sich die Entwickler von Deepfakes darauf einstellen – der Fortschritt kommt also immer beiden Seiten zugute. Aus diesem Grund müssen jene Plattformen stärker in die Pflicht genommen werden, über die die manipulierten Inhalte verbreitet werden. Zwar haben sich Facebook und Twitter mittlerweile selbst Regeln zum Umgang mit manipulierten Inhalten auferlegt, diese sind aber einerseits nicht einheitlich, anderseits ist es nicht wünschenswert, die Definition der Meinungsfreiheit privaten Unternehmen zu überlassen.

Die Bundesregierung ist auf das Thema „Einsatz KI-manipulierter Inhalte zur Desinformation“ nicht vorbereitet, wie eine kleine Anfrage der FDP Bundestagsfraktion vom Dezember 2019 deutlich gezeigt hat: Es gibt keine klare Zuständigkeit für das Thema und auch keine spezifische Gesetzgebung; bislang finden lediglich „generell-abstrakte Regelungen“ Anwendung. Aus den Antworten der Bundesregierung lassen sich weder eine konkrete Strategie noch eine Investitionsabsicht ableiten.

Insgesamt scheinen die bisherigen Regulierungsversuche auf deutscher und europäischer Ebene kaum geeignet, KI-basierte Desinformation einzudämmen. Dabei ginge es auch anders. In einigen US-Bundesstaaten gibt es bereits Gesetze sowohl gegen nicht einvernehmliche Deepfake-Pornografie

EXECUTIVE SUMMARY

als auch gegen den Einsatz dieser Technologie zur Beeinflussung von Wählern. Vor diesem Hintergrund sollte der Gesetzgeber klare Richtlinien für den einheitlichen Umgang der digitalen Plattformen mit Deepfakes im Speziellen und mit Desinformation im Allgemeinen schaffen. Die Maßnahmen können von der Kennzeichnung manipulierter Inhalte über die Limitierung ihrer Verbreitung (Ausschluss aus Empfehlungsalgorithmen) bis zu ihrer Löschung reichen. Zudem sollte die Förderung von Medienkompetenz für alle Bürgerinnen und Bürger, unabhängig vom Alter, eine Priorität darstellen. Es ist wichtig, die Öffentlichkeit für die Existenz von Deepfakes zu sensibilisieren und die Kompetenz der Menschen zu fördern, audiovisuelle Inhalte zu hinterfragen – auch wenn es immer schwieriger wird, sie als Fälschungen zu erkennen. In dieser Hinsicht lohnt ein Blick auf die nordischen Länder, insbesondere auf Finnland, dessen Bevölkerung die höchste Widerstandsfähigkeit gegenüber Desinformation aufweist.

Eines sollten wir allerdings nicht tun: der Versuchung nachgeben, Deepfakes grundsätzlich zu verbieten. Wie jede Technologie, so eröffnet auch diese jenseits der mit ihr verbundenen Gefahren eine Fülle von interessanten Möglichkeiten – unter anderem für Bildung, Film und Satire.

1. STAND DER ENTWICKLUNG

Künstliche Intelligenz und ihre Rolle bei der Desinformation

Obwohl die Wurzeln der Künstlichen Intelligenz bis in die Mitte des 20. Jahrhunderts zurückreichen, hat diese Technologie lange Zeit wenig Aufmerksamkeit erfahren. Erst mit Beginn der 2010er Jahre scheint das Ende des langen KI-Winters eingeläutet worden zu sein: Im Jahr 2011 schlug IBMs Computersystem Watson die besten menschlichen Spieler in der Fernseh-Quizshow Jeopardy [1]), Googles selbstfahrender Autoprototyp legte mehr als 100.000 Meilen (160.000 Kilometer) zurück, und Apple stellte die „intelligente persönliche Assistenz" Siri vor. Das öffentliche Interesse an Künstlicher Intelligenz, vor allem an den mit ihr verbundenen Risiken, ist seither stetig gewachsen. Der Diskurs über Superintelligenz – ausgelöst durch das gleichnamige Buch von Nick Bostrom, das 2014 erschienen ist – hat die Aufmerksamkeit noch weiter gesteigert. Seitdem haben sich prominente Persönlichkeiten immer wieder mahnend, teils alarmierend zu dem Thema geäußert. Häufig zitiert werden Stephen Hawking („Die Entwicklung der künstlichen Intelligenz könnte das Ende der Menschheit bedeuten") und Elon Musk („KI ist ein grundlegendes existenzielles Risiko für die menschliche Zivilisation"). Während Superintelligenz und auch die sogenannte „starke KI" (AGI, Artificial General Intelligence) noch in ferner Zukunft liegen, spielt die „schwache KI" mit ihren gar nicht so schwachen Algorithmen bereits heute eine immer größer werdende Rolle in Wirtschaft, Gesellschaft und Politik. Die Autorin ist davon überzeugt, dass die Auswirkungen auf Gesundheit, Energie, Sicherheit, Mobilität und viele weitere Bereiche weitgehend positiv sein werden. Wir werden aber die positiven Seiten der Entwicklungen nur dann genießen können, wenn wir die Risiken dieser Technologie erkennen und ihnen erfolgreich entgegenwirken. Eines dieser Risiken ist der Missbrauch der Technologie für die bewusste Verbreitung von falschen Informationen. Politisch motivierte Desinformation ist selbstverständlich kein neues Phänomen. Stalin und Mao sind die prominentesten Namen unter jenen Diktatoren, die regelmäßig Fotografien so bearbeiten ließen, dass alte Bilder

jeopardy watson (c) Kaelson-Jeopardy Productions Inc., via Associated Press

mit der aktuellen „Wahrheit“ übereinstimmten: Wer nicht länger genehm war, wurde aus den Bildern gelöscht, wer neu in die Parteispitze kam, wurde nachträglich hinzugefügt; auch der Kontext der Bilder wurde, zum Beispiel durch einen abweichenden Hintergrund, verändert.

> *„Wir werden die positiven Aspekte dieser Technologie nur dann geniessen können, wenn wir ihre Risiken erkennen und ihnen erfolgreich entgegenwirken.“*

Mit der manipulierten visuellen Aufzeichnung sollten neue Fakten geschaffen, Geschichten und Geschichte neu geschrieben werden. Damals waren solche Anpassungen langwierig und setzten Spezialwissen voraus, heute kann dies – mit der richtigen App auf dem Smartphone – jede und jeder problemlos selbst erledigen. Und bei Fotos macht die Technologie nicht halt. Ein gefälschtes Video zu produzieren, das glaubwürdig aussieht, ist aktuell zwar noch mit einigem Aufwand verbunden. Doch durch bestimmte Verfahren der Künstlichen Intelligenz wird es zusehends einfacher, existierende Videos zu manipulieren. Diese Videos sind mittlerweile unter dem Namen Deepfakes bekannt geworden. Noch sind sie im Internet selten zu finden, doch mit zunehmender Nutzung und Verbreitung stellen sie eine immer größer werdende Herausforderung für unsere Gesellschaft dar.

Manipulierte Inhalte verbreiten sich auf Plattformen wie Facebook oder YouTube nicht nur in hoher Geschwindigkeit, sie werden aufnahmewilligen Empfängerinnen und Empfängern auch gezielt angezeigt. Zudem verlagert sich die Verbreitung von Desinformation zunehmend auf Messengerdienste wie zum Beispiel WhatsApp.

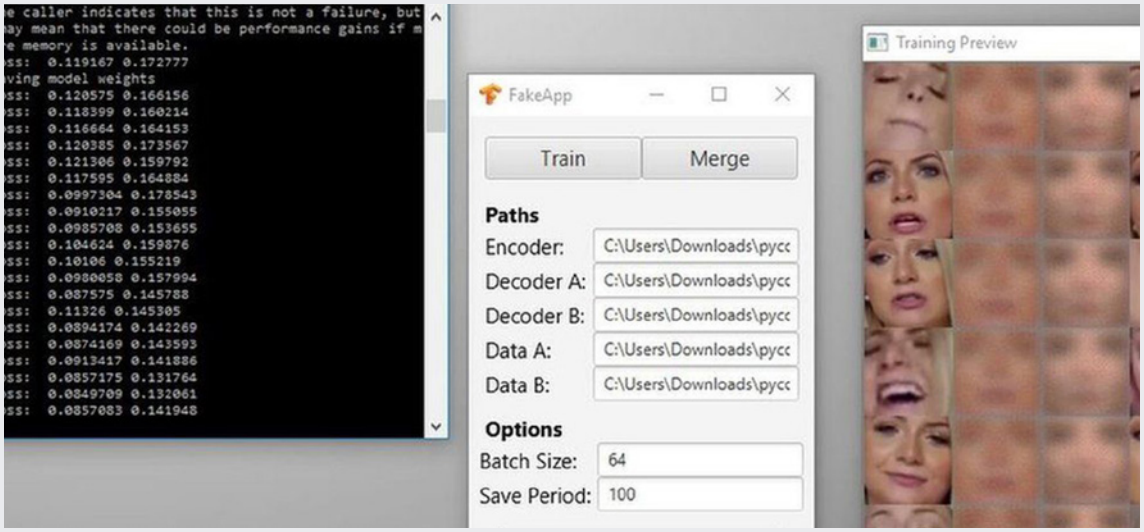
Dort werden verschlüsselte Nachrichten über private Verbindungen verbreitet, was das Vertrauen in die weitergeleiteten Informationen erhöht; es wird dadurch eine Art versteckte Viralität erzeugt. Die Verschlüsselung der privaten Online-Kommunikation ist, ähnlich wie das Briefgeheimnis, ein erstrebenswertes Gut – auf diese Weise können Nachrichten nicht von Dritten eingesehen werden. Die Verschlüsselung bedeutet aber auch, dass die dort verbreiteten Informationen nicht auf ihren Wahrheitsgehalt überprüft und somit nicht entsprechend moderiert werden können.

2. CHEAP FAKES & DEEPFAKES

Technologische Möglichkeiten der Manipulation von Text, Bild, Audio und Video

In den vergangenen zwei Jahren hat der Begriff Deepfake konstant an Bekanntheit hinzugewonnen. Doch was genau sind Deepfakes und wie unterscheiden sie sich von anderen manipulierten Inhalten? Während die ersten wissenschaftlichen KI-Experimente zur Manipulation von Videos bereits Ende der 1990er Jahre erfolgten, erfuhr die breite Öffentlichkeit erst ab Ende 2017 von dieser technischen Möglichkeit. Zu diesem Zeitpunkt ist auch die Begrifflichkeit entstanden, als ein Reddit-Benutzer namens Deepfakes und andere Mitglieder der Reddit-Community „r/deepfakes“ die von ihnen erstellten Inhalte veröffentlichten.

> *„Bei den ersten Deepfakes handelte es sich, nicht sehr überraschend, um pornografische Videos, in denen die Gesichter der Darstellerinnen durch die von Prominenten wie Scarlett Johansson oder Taylor Swift ersetzt wurden.“*



Fake App Screenshot (c) BBC UK

In vielen Fällen handelte es sich, nicht sehr überraschend, um pornografische Videos, in denen die Gesichter der Darstellerinnen durch die von Prominenten wie Scarlett Johansson oder Taylor Swift ersetzt wurden. Zu den etwas harmloseren Beispielen zählten Filmszenen, in denen alle Gesichter der Schauspielerinnen und Schauspieler gegen das von Nicolas Cage getauscht wurden.

2.1 Deepfakes vs Cheap Fakes

Auch wenn die Manipulation von Pornografie mit Sicherheit zu den am meisten verbreiteten Beispielen von Deepfakes gehört, ist sie nicht der Hauptgrund für die aktuelle gesellschaftliche Debatte. Interessanterweise war das Video, das diese Debatte angestoßen hat, überhaupt kein Deepfake, sondern ein Cheap Fake (manchmal auch Shallow Fake genannt): ein mit sehr einfachen technischen Mitteln gefälschtes Video von der Sprecherin des US-Repräsentantenhauses, Nancy Pelosi. Die Originalgeschwindigkeit der Aufnahme wurde auf etwa 75 Prozent reduziert und die Tonhöhe angehoben, um den natürlichen Klang der Stimme zu erhalten.

Ergebnis: Wer das Video betrachtete, konnte den plausiblen Eindruck gewinnen, dass Nancy Pelosi betrunken war. Es wurde millionenfach in den sozialen Medien geteilt. Dies zeigt, wie schon einfachste Fälschungen die Realität verzerren und

Pelosi Deepfake (c) CBS

zu politischen Zwecken eingesetzt werden können. Immerhin war es bislang sehr schwierig, die Aufnahme dahingehend zu verfälschen, dass die betroffene Person ganz andere Bewegungen vorführt oder ganz andere Worte ausspricht als im Originalvideo. Bislang.

WIE FUNKTIONIEREN EIGENTLICH DEEPFAKES?

Deepfakes (eine Wortverschmelzung von Deep Learning und Fake, englisch für Fälschung) sind das Produkt zweier KI-Algorithmen, die in einem sogenannten Generative Adversarial Network (zu Deutsch „erzeugenden gegnerischen Netzwerk"), abgekürzt GAN, zusammenarbeiten.

Die GANs können am besten als eine Möglichkeit beschrieben werden, algorithmisch neue Arten von Daten aus bestehenden Datensätzen zu generieren. So könnte ein GAN beispielsweise Tausende von Aufnahmen von Donald Trump analysieren und dann ein neues Bild erstellen, das den ausgewerteten Aufnahmen ähnelt, ohne aber eine exakte Kopie einer dieser Aufnahmen zu sein. Diese Technologie kann auf unterschiedliche Arten von Inhalten – Bild, Bewegtbild, Ton und Text – angewendet werden. Die Bezeichnung Deepfake wird aber vor allem auf Audio- und Videoinhalte angewendet.

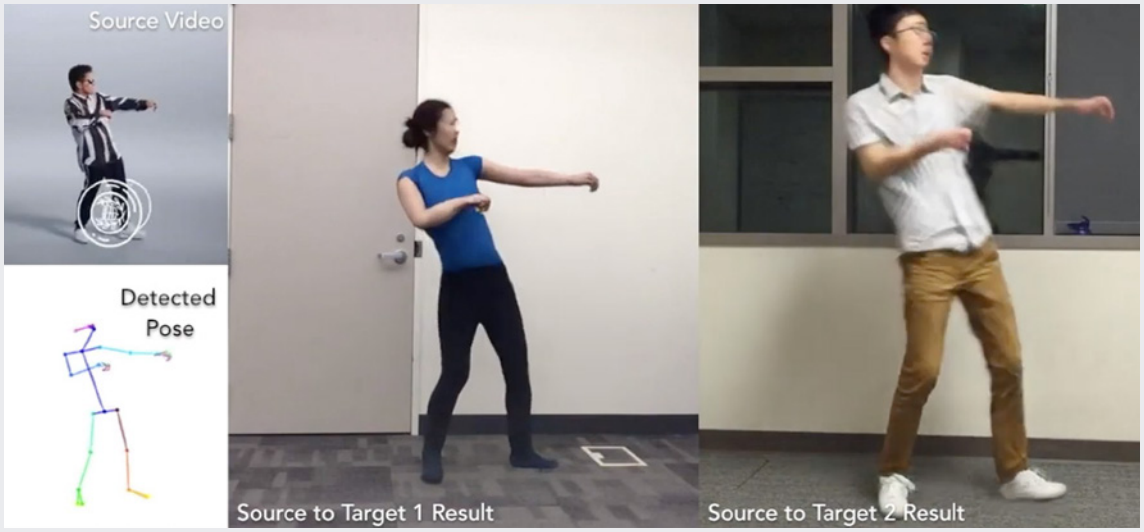
Mittlerweile sind für ein glaubwürdiges Ergebnis Trainingsdaten von nur wenigen Hundert Bildern bzw. Tonaufnahmen erforderlich. Schon für knappe 3 US-Dollar kann jeder ein gefälschtes Video einer beliebigen Person bestellen, vorausgesetzt, es stehen mindestens 250 Bilder dieser Person zur Verfügung – das dürfte aber bei den meisten Personen, die Instagram oder Facebook nutzen, kein Problem sein. Auch synthetische Sprachaufnahmen lassen sich für lediglich 10 US-Dollar per 50 Wörter generieren.

2.2 Beispiele der Anwendung

Manipulation von Bewegungsmustern

Große Aufmerksamkeit hat 2018 eine Anwendung von vier Berkeley-Forschern erhalten, die Künstliche Intelligenz verwendet, um die Tanzschritte einer Ausgangsperson (zum Beispiel einer professionellen Tänzerin) auf eine Zielperson zu übertragen[2]).

Ausgehend vom Quellvideo werden die Bewegungen auf ein „Strichmännchen" übertragen. Im nächsten Schritt synthetisiert das neuronale Netzwerk das Zielvideo gemäß den „Strichmännchenbewegungen". Das Ergebnis ist ein „gefaktes" Video, in dem eine dritte Person wie ein Profi tanzt. Natürlich kann ein solcher Algorithmus nicht nur dafür verwendet werden, Tanzbewegungen zu imitieren, sondern potenziell jede andere Form von Bewegung.



Dance Now (c) UC Berkley

Damit sind Tür und Tor geöffnet, um politische Gegner in kompromittierenden Situationen darzustellen. Welche Auswirkung hätte zum Beispiel eine Videoaufnahme, die einen Politiker mit Hitlergruß zeigt oder einfach nur beim Zeigen des Stinkefingers?

Stimme und Mimik

Noch weiter reichende Konsequenzen können Fälschungen haben, in denen Personen Worte in den Mund gelegt wurden, die sie nie gesagt haben, in denen aber Gestik, Mimik und Stimme verblüffend echt wirken. Mehrere solcher Videos, unter anderem von Barack Obama und Mark Zuckerberg, wurden erstellt, allerdings nicht um die Zuschauer zu täuschen, sondern um die Möglichkeiten der Technologie und ihre Gefahren zu demonstrieren. Inzwischen wurde ein Deepfake auch von einer politischen Partei, der belgischen Socialistische Partij Anders (sp.a), erstellt und verbreitet.

Im Mai 2018 hat sie ein Video, in dem Trump Belgien dafür verspottet, dass es dem Pariser Klimaabkommen treu bleibt, auf Facebook gepostet[3]. Trotz der offensichtlich schlechten Qualität und einer eher unnatürlichen Mundbewegung, die einen aufmerksamen Zuschauer sofort Verdacht schöpfen lassen sollte, provozierte es Hunderte von Kommentaren, in denen viele ihre Empörung darüber zum Ausdruck brachten, dass der amerikanische Präsident es wagen würde, sich in die belgische Klimapolitik einzumischen.

Auch im Falle dieses Videos ging es den Machern um Aufklärung. Das Video war eine gezielte Provokation, um die Aufmerksamkeit der Menschen auf eine Online- Petition zu lenken, in der die belgische Regierung zu dringenden Klimaschutzmaßnahmen aufgefordert wird.

Was wäre aber, wenn jemand ein Video erstellen würde, in dem Trump nicht über die belgische Klimapolitik spricht, sondern zum Beispiel darüber, dass er einen Atomangriff auf Iran beabsichtigt?

Bildmanipulation: DeepNude und künstliche Gesichter

Inhalte, die häufig nicht zu den Deepfakes gezählt werden, obwohl sie mit einer sehr ähnlichen Technologie generiert werden, sind Bild- und Textinhalte. Der Grund dafür ist einfach: Sowohl Bilder als auch Texte können ohne den Einsatz komplexer Technologie so leicht manipuliert werden, dass der „Mehrwert“ (oder der Nachteil, je nach Perspektive) im Vergleich zu Audio- und Videoinhalten gering ausfällt. Außerdem sind

Ein mit der Anwendung deepnude.to generiertes Nacktbild

Deepfake_Example (c) Agnieszka Walorska

WAS SIND KÜNSTLICHE NEURONALE NETZE?

Künstliche neuronale Netze (= Artificial Neural Networks, kurz ANN) sind Computersysteme, die vage von biologischen neuronalen Netzen inspiriert sind, welche sich in den Gehirnen von Menschen und Tieren befinden.

ANN „lernen“ die Ausführung von Aufgaben anhand von Beispielen, ohne mit aufgabenspezifischen Regeln programmiert zu sein. Sie können zum Beispiel lernen, Bilder zu identifizieren, die Katzen enthalten, indem sie Beispielbilder analysieren, die manuell als „Katze“ oder „keine Katze“ gekennzeichnet wurden, und die Ergebnisse zur Identifizierung von Katzen in anderen Bildern verwenden

baldwin trump (c) NBCU

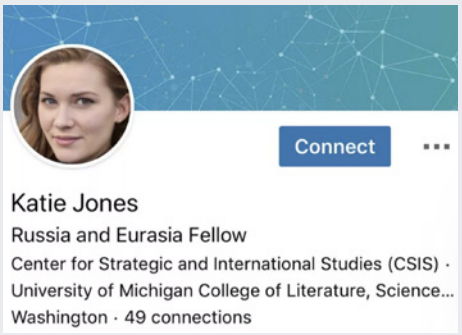
Mit thispersondoesnotexist.com zufällig generierte Gesichter

Deepfakes-Example (c) thispersondoesnotexist.com

Videoaufnahmen im Vergleich zu Text und statischem Bild viel effektiver, um Emotionen wie Angst, Wut oder Hass auszulösen.

Nichtsdestotrotz haben einige der Beispiele für KI-basierte Manipulationen solcher Inhalte für Aufmerksamkeit gesorgt. Wie schon bei Videos, so werden auch bei Bildern die Algorithmen vor allem dazu verwendet, gefälschte pornografische Inhalte zu erzeugen. Innerhalb weniger Sekunden können Anwendungen wie DeepNude ein Bikinifoto in ein sehr realistisches Nacktbild umwandeln. Es wird wohl niemanden überraschen, dass die App nur bei Frauen funktioniert (beim Versuch, das Bild eines Mannes zu verwenden, werden einfach weibliche Genitalien generiert) und damit jede Frau zu einem potenziellen Opfer von „Rachepornos“ (Revenge Porn) macht, auch wenn kein einziges echtes Nacktbild von ihr existiert.

Die neuronalen Netze können im Übrigen nicht nur zur Manipulation von Bildern existierender Personen angewendet werden, sie „erschaffen“ auch ganz neue Personen – oder zumindest ganz neue Gesichter. Eine kommerzielle Anwendung für diese Technologie liegt auf der Hand: Bilddatenbanken können mit KI deutlich kosteneffizienter bestückt werden als mit dem Einsatz von echten Menschen. Allerdings bedeutet dies auch, dass die Erstellung von falschen Social-Media-Profilen, die zum Beispiel zur Verbreitung bestimmter politischer Inhalte eingesetzt werden können, deutlich erleichtert wird. Auch Spionageversuche mit computergenerierten Profilbildern werden bereits vermutet, zum Beispiel bei einem LinkedIn-Profil von „Katie Jones“, einer angeblichen Forscherin in einem US-amerikanischen Think-Tank.



LinkedIn Profil: AI-generiertes Profil für “Katie Jones“

KI Profile Screenshot (c) LinkedIn, AP

Bevor eine Expertenanalyse mehrere visuelle Anomalien identifizierte, die darauf hindeuteten, dass das Bild synthetisch erzeugt wurde, hat das Profil es geschafft, sich mit 52 politischen Persönlichkeiten in Washington zu verknüpfen, darunter einem stellvertretenden Assistant Secretary of State, einem hochrangigen Berater eines Senators und einem prominenten Wirtschaftswissenschaftler[4].

Das Konto wurde von LinkedIn schnell entfernt, soll aber zu einem Netzwerk von Phantomprofilen gehören, von denen einige möglicherweise weiterhin existieren und beispielsweise für Phishing-Attacken eingesetzt werden können.

KI-generierte Texte

Die beschriebene Anwendung kann sich besonders dann entfalten, wenn sie mit den Mitteln verknüpft wird, die eine KI-getriebene Textgenerierung bietet. Viele haben von dieser Möglichkeit im Kontext des von dem Forschungsunternehmen OpenAI geschaffenen Textgenerators GPT-2 gehört, der wegen seines Missbrauchspotenzials ursprünglich als zu gefährlich angesehen wurde, um ihn der Öffentlichkeit zur Verfügung zu stellen[5].

Später hat sich das Unternehmen doch dazu entschlossen, GPT-2 in mehreren Schritten zu veröffentlichen, da die Macherinnen und Macher bis dato keine eindeutigen Beweise für einen Missbrauch feststellen konnten[6]. Obwohl dies tatsächlich bislang der Fall sein mag, räumen sie gleichzeitig ein, dass die Menschen die vom GPT-2 generierten Texte zum größten Teil für glaubwürdig erachten würden, dass der Generator für extremistische Inhalte feinjustiert werden könne und dass die Erkennung der generierten Texte eine Herausforderung darstelle.

AI-generated fake content could unleash a virtual arms race of misinformation online, experts say.

> **Once you get the person to click something, you've gotten them to put themselves in a position to think a certain way, if they haven't already done so", said Katherine Jellison, a professor at Georgia Tech's School of Interactive Computing and author of the book "Cyberbullying in the Age of the Internet."**

Hervorgehobener Text – Vorgabe, übriger Text – KI-generiert mit www.talktotransformer.de

Mit der Anwendung „Talk To Transformer" kann jede und jeder die Funktionsweise von GPT-2 ausprobieren. Gibt man in den Generator einen oder mehrere Sätze ein, erzeugt er einen Text, der die Eingabe als Ausgangspunkt nimmt. Die Ergebnisse sind oft – nicht immer – überraschend kohärent. Sie treffen den zur Vorgabe passenden Ton und simulieren Glaubwürdigkeit mit erfundenen Experten, Statistiken und Zitaten.

3. VERBREITUNG & KONSEQUENZEN

Wir gefährlich sind Deepfakes wirklich?

3.1 Verbreitung

Es ist nicht einfach, die Verbreitung von Deepfakes exakt zu quantifizieren, zumal davon ausgegangen werden kann, dass ihre Anzahl stetig wächst. Das Unternehmen Deeptrace, das eine technologische Lösung für die Erkennung von Deepfakes anbietet, hat sich in seinem Report „The State of Deepfakes: Landscape, Threats, and Impact"[7] um eine genaue Schätzung bemüht. Dem im September 2019 veröffentlichten Bericht zufolge hat sich die Zahl von Deepfakes innerhalb von sieben Monaten von 7.964 im Dezember 2018 auf 14.678 im Juli 2019 fast verdoppelt. Bei 96 Prozent dieser Deepfakes handelt es sich um nicht einvernehmlich erzeugte pornografische Inhalte, die ausschließlich weibliche Körper zeigen.

Vorrangig betroffen sind prominente Frauen, deren gefälschte Bilder zu Tausenden online verfügbar sind. Allein die vier populärsten DeepPorn-Websites verzeichnen laut Deeptrace-Report inzwischen mehr als 134 Millionen Aufrufe gefälschter Videos von weiblichen Prominenten. Aber auch viele Privatpersonen sind von der bereits erwähnten Rachepornografie betroffen. Der Anstieg wird vor allem durch die bessere Zugänglichkeit sowohl von Werkzeugen als auch von Dienstleistungen ermöglicht, die die Erstellung von Deepfakes auch ohne Programmierkenntnisse möglich machen.

Im Jahr 2019 wurde auch bereits über Fälle berichtet, in denen KI-generierte Sprachklone für Social Engineering verwendet wurden. Im August berichtete The Wall Street Journal[8] von einem ersten Fall KI-basierten Stimmbetrugs – auch bekannt

als Vishing (kurz für Voice Phishing) –, der das betroffene deutsche Unternehmen 220.000 Euro kostete.

Die Software hat die Stimme des deutschen Managers, samt Melodie und dem leichten deutschen Akzent, so erfolgreich nachgeahmt, dass sein britischer Kollege sofort dem dringenden Wunsch des Anrufers nachgekommen ist, die genannte Summe zu überweisen. Es handelt sich zwar bisher um einen Einzelfall, doch ist davon auszugehen, dass es solche Versuche in Zukunft häufiger geben wird.

Ein signifikanter Teil der medialen Berichterstattung über Deepfakes hat sich auf ihr Potenzial konzentriert, politische Gegner zu diskreditieren und demokratische Prozesse zu untergraben. Dieses Potenzial hat sich bisher nicht entfaltet. Zwar wurden Videos von Politikern wie Barack Obama, Donald Trump oder Matteo Renzi technisch manipuliert, dies geschah bisher allerdings primär zu Satire- oder Demonstrationszwecken und wurde schnell aufgeklärt.

3.2 Konsequenzen

Die Tatsache, dass bisher keine Deepfakes von Politikern zur Desinformation verwendet wurden, bedeutet jedoch mitnichten, dass sie keine Auswirkungen auf den politischen Diskurs hatten. Ein Beispiel, das in den westlichen Medien nur wenig Beachtung fand, zeigt, wie das bloße Wissen um die Existenz von Deepfakes das politische Klima beeinflussen kann:

Gabuns Präsident, Ali Bongo, hatte nach einem Schlaganfall monatelang keine öffentlichen Auftritte absolviert. Nachvollziehbarerweise kochte die Gerüchteküche hoch, und es wurden Stimmen laut, die behaupteten, der Präsident sei verstorben. Um die Spekulationen auszuräumen, wurde im Dezember 2018 ein Video veröffentlicht, in dem er seine übliche Neujahrsansprache hielt. Die Aufnahme hatte jedoch einen konträren Effekt. Viele waren der Meinung, Bongo hätte seltsam ausgesehen, und vermuteten sofort, dass es sich bei dem Video um eine Fälschung handele. Kurz darauf startete das Militär einen missglückten Staatsstreich und nannte den vermeintlichen Deepfake als Teil der Motivation[9].

Die anschließend vorgenommene forensische Analyse hat allerdings die Echtheit der Aufnahme bestätigt. Ali Bongo hat sich inzwischen von seinem Schlaganfall erholt und ist weiterhin im Amt. Dies zeigt, dass die größte Bedrohung durch Deepfakes gar nicht die Deepfakes selbst sein müssen. Allein die technologische Möglichkeit, solche Videos zu erstellen, wirft die Frage auf:

> *„Obwohl es technisch manipulierte Videos von Politkern wie Barack Obama, Donald Trump und Matteo Renzi gegeben hat, waren sie in erster Linie satirisch motiviert oder zu Demonstrationszwecken erstellt worden, und ihre Fälschungen wurden schnell aufgedeckt.“*

Kann man der Authentizität von Bewegtbildern noch vertrauen? Diese Frage wirft ihre Schatten auf die im Jahr 2020 stattfindenden US-Präsidentschaftswahlen. Bereits im Wahlkampf 2016 haben KI-gestützte Desinformation und Manipulation, vor allem in Form von Microtargeting und Bots, eine Rolle gespielt. Mit Deepfakes ist nun ein weiteres Instrument zum Desinformationsarsenal hinzugekommen. Auch wenn keine oder nur wenige tatsächliche Deepfakes in dem Wahlkampf angewendet werden sollten, werden Politikerinnen und Politiker höchstwahrscheinlich die Möglichkeit dankbar annehmen, echte, aber unvorteilhafte Aufnahmen als solche abzutun.

3.3 Gibt es auch positive Beispiele für die Anwendung von Deepfakes?

„Die Technologie gibt uns [...] Wege, Schaden anzurichten und Gutes zu tun; sie verstärkt beides. [...] Aber die Tatsache, dass wir auch jedes Mal eine neue Wahl haben, ist ein neues Gut“[10], wird Kevin Kelly, der langjährige Chefredakteur und Mitglied des Gründungsteams des Technologiemagazins Wired, zitiert. Kann diese Aussage auch auf Deepfakes zutreffen?

> *„Die Technologie ist besonders vielversprechend für die Film-Industrie, insbesondere in der Postproduktion und der Synchronisation."*

Interessant ist die Technologie besonders für die Filmbranche und hier vor allem in der Postproduktion und Synchronisation. Warum? Gegenwärtig müssen die Filmstudios einen großen Aufwand betreiben, um einen Dialog nachträglich anzupassen. Die beteiligten Schauspielerinnen und Schauspieler, das notwendige Personal und der Drehort müssen in einem solchen Fall noch einmal gebucht werden. Mit der Technologie, die den Deepfakes zugrunde liegt, könnte man solche Veränderungen innerhalb kürzester Zeit und zu einem Bruchteil der Kosten durchführen. Auch die Synchronisation der Filme könnte deutlich verbessert werden:

Es wäre möglich, die Lippenbewegungen der Schauspielerinnen und Schauspieler an die Worte der Synchronsprecherinnen und Synchronsprecher anzupassen oder die Stimmen gleich zu synthetisieren und an die entsprechende Sprache anzupassen, so dass keine Synchronisation mehr notwendig ist. Ein Beispiel für solch einen Einsatz liefert ein Video von David Beckham, der für eine Kampagne gegen Malaria wirbt[11]. Er „spricht" darin in mehreren Sprachen – und jedes Mal scheint sein Mund mit den Worten perfekt synchronisiert zu sein.

Auch der Bereich Bildung stellt ein interessantes Einsatzgebiet dar: So könnten beispielsweise Videos von historischen Figuren erstellt werden, die ihre Geschichte erzählen oder Fragen beantworten. Für viel Medienecho hat das Projekt „Dimensions of History"[12] der Shoah Foundation der University of Southern California gesorgt, bei dem Interviews mit 15 Holocaust-Überlebenden geführt und holografische Aufnahmen von ihnen gemacht wurden.

Die Wanderausstellung war in verschiedenen Museen in den USA und zuletzt auch im schwedischen Historischen Museum zu sehen. Die Besucherinnen und Besucher der Ausstellung hatten anschließend die Möglichkeit, ihre Fragen an die Hologramme zu stellen. Die Spracherkennungssoftware ordnete die Frage einem Interviewausschnitt zu. Mit dem Einsatz der Deepfake- Technologie könnte dasselbe in größerem Maßstab und mehrsprachig durchgeführt werden.

4. DEEPFAKES BEKÄMPFEN

Wie können wir den mit Deepfakes verbundenen Herausforderungen begegnen?

Diese positiven Beispiele sollen selbstverständlich nicht die potenziellen Gefahren, die von Deepfakes ausgehen, kleinreden. Diese Gefahren sind unbestritten und sie erfordern entschiedene Gegenmaßnahmen – darüber besteht weitgehend Einigkeit. Weniger Einigkeit besteht darüber, wie genau diese Gegenmaßnahmen auszusehen haben. Es stellt sich zudem die Frage, wie das Recht des Einzelnen auf freie Meinungsäußerung garantiert werden kann, ohne dass gleichzeitig das Bedürfnis der Gesellschaft nach einem zuverlässigen Informationssystem untergraben wird.

4.1 Technische Lösungen zur Identifikation und Bekämpfung von Deepfakes

Eine Möglichkeit, gegen die Fälschungen vorzugehen, besteht darin, Technologien zu entwickeln, die Fälschungen von realen Inhalten unterscheiden können. Zu diesem Zweck werden Algorithmen verwendet, die jenen ähneln, die zur Erzeugung von Täuschungen entwickelt wurden. Mit GLTR, einem auf dem bereits erwähnten GPT-2 basierenden Modell, untersuchten Forscher des MIT-IBM Watson AI Lab und der HarvardNLP, ob dieselbe Technologie, die eigenständig erfundene Artikel schreibt, auch dafür genutzt werden kann, durch KI generierte Passagen zu erkennen.

Gibt man in die Testanwendung eine Textpassage ein, hebt sie die Wörter in Grün, Gelb, Rot oder Violett hervor, um die abnehmende Vorhersagbarkeit anzuzeigen. Je höher der Anteil an Wörtern mit geringer Vorhersagbarkeit, also an rot- und violettmarkierten Sätzen, desto größer ist die Wahrscheinlichkeit, dass es sich bei der Passage um den Text eines menschlichen Autors oder einer menschlichen Autorin handelt; je vorhersagbarer die Wörter (und „grüner“ die Passagen), desto wahrscheinlicher ist es wiederum, dass es sich um einen Textgenerator handelt.

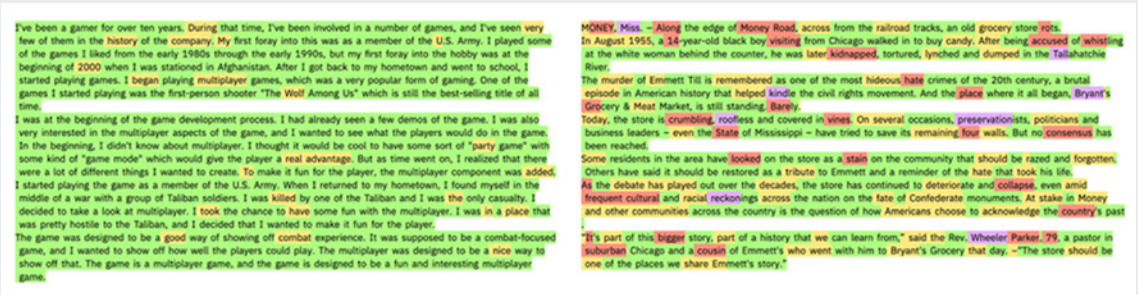
Ähnliche Verfahren können angewendet werden, um manipulierte Videos zu enttarnen. Im Jahr 2018 stellten Forscher fest, dass die Gesichter in Deepfake-Videos nicht blinzeln. Dies lag daran, dass statische Bilder zur Generierung der Aufnahmen genutzt wurden und diese meist Menschen mit offenen Augen zeigten. Doch der Nutzen dieser Erkenntnis hatte keine lange Dauer. Sobald diese Information publik wurde, tauchten die ersten Videos mit blinzelnden Augen auf. Ähnlich wird es sich in Zukunft mit anderen Entdeckungsmechanismen verhalten. Das gleiche Katz-und-Maus-Spiel wird im Bereich der Cybersecurity mittlerweile seit Jahrzehnten gespielt – der Fortschritt kommt immer beiden Seiten zugute.

Diese Tatsache ist freilich kein Grund, Anstrengungen zur Deepfake-Identifizierung ruhen zu lassen. Im September 2019 kündigte Facebook – in Zusammenarbeit mit der PAI-Initiative[13], Microsoft und mehreren Universitäten – eine mit 10 Millionen US- Dollar dotierte „Deepfake Detection Challenge“[14] an.

Facebook hat auch die Erstellung eines Datensatzes mit Bildern und Videos von zu diesem Zweck engagierten Schauspielern in Auftrag gegeben, um eine ausreichende Datengrundlage für die Challenge zu schaffen. Wenige Wochen später veröffentlichte auch Google einen Datensatz von 3.000 manipulierten Videos mit dem gleichen Ziel.

Auch die beim Pentagon angesiedelte US-amerikanische Forschungsförderungsagentur DARPA arbeitet bereits seit 2016 innerhalb des MediFor-Programms (kurz für Media Forensics) daran, manipulierte Inhalte zu erkennen, und hat dafür innerhalb von zwei Jahren 68 Millionen US-Dollar investiert.[15] Ob und an welchen technischen Lösungen zur Bekämpfung von Deepfakes in Deutschland und Europa gearbeitet wird, ist wenig bekannt.

Meist handelt es sich hierbei um einzelne Unternehmen, wie das bereits erwähnte Deeptrace, und um Forschungsprojekte


I've been a gamer for over ten years. During that time, I've been involved in a number of games, and I've seen very few of them in the history of the company. My first foray into this was as a member of the U.S. Army. I played some of the games I liked from the early 1980s through the early 1990s, but my first foray into the hobby was at the beginning of 2000 when I was stationed in Afghanistan. After I got back to my hometown and went to school, I started playing games. I began playing multiplayer games, which was a very popular form of gaming. One of the games I started playing was the first-person shooter "The Wolf Among Us" which is still the best-selling title of all time.
I was at the beginning of the game development process. I had already seen a few demos of the game. I was also very interested in the multiplayer aspects of the game, and I wanted to see what the players would do in the game. In the beginning, I didn't know about multiplayer. I thought it would be cool to have some sort of "party game" with some kind of "game mode" which would give the player a real advantage. But as time went on, I realized that there were a lot of different things I wanted to create. To make it fun for the player, the multiplayer component was added.
I started playing the game as a member of the U.S. Army. When I returned to my hometown, I found myself in the middle of a war with a group of Taliban soldiers. I was killed by one of the Taliban and I was the only casualty. I decided to take a look at multiplayer. I took the chance to have some fun with the multiplayer. I was in a place that was pretty hostile to the Taliban, and I decided that I wanted to make it fun for the player.
The game was designed to be a good way of showing off combat experience. It was supposed to be a combat-focused game, and I wanted to show off how well the players could play. The multiplayer was designed to be a nice way to show off that. The game is a multiplayer game, and the game is designed to be a fun and interesting multiplayer game.

MONEY, Miss. – Along the edge of Money Road, across from the railroad tracks, an old grocery store rots.
In August 1955, a 14-year-old black boy visiting from Chicago walked in to buy candy. After being accused of whistling at the white woman behind the counter, he was later kidnapped, tortured, lynched and dumped in the Tallahatchie River.
The murder of Emmett Till is remembered as one of the most hideous hate crimes of the 20th century, a brutal episode in American history that helped kindle the civil rights movement. And the place where it all began, Bryant's Grocery & Meat Market, is still standing. Barely.
Today, the store is crumbling, roofless and covered in vines. On several occasions, preservationists, politicians and business leaders – even the State of Mississippi – have tried to save its remaining four walls. But no consensus has been reached.
Some residents in the area have looked on the store as a stain on the community that should be razed and forgotten. Others have said it should be restored as a tribute to Emmett and a reminder of the hate that took his life.
As the debate has played out over the decades, the store has continued to deteriorate and collapse, even amid frequent cultural and racial reckonings across the nation on the fate of Confederate monuments. At stake in Money and other communities across the country is the question of how Americans choose to acknowledge the country's past .
"It's part of this bigger story, part of a history that we can learn from," said the Rev. Wheeler Parker, 79, a pastor in suburban Chicago and a cousin of Emmett's who went with him to Bryant's Grocery that day. –"The store should be one of the places we share Emmett's story."

Analyseergebnis: menchlicher Autor vs. Textgenerator

Deepfakes-Example-3 (c) gltr.io

> *„Je höher der Anteil an Wörtern mit geringer Vorhersagbarkeit, also an rot- und violettmarkierten Sätzen, desto größer ist die Wahrscheinlichkeit, dass es bei der Passage um den Text eines menschlichen Autors oder einer menschlichen Autorin handelt; je vorhersagbarer die Wörter (und "grüner" die Passagen), desto wahrscheinlicher ist es wiederum, dass es sich um einen Textgenerator handelt."*

wie Face2Face[16] von Matthias Nießner, Professor an der TU München. Laut Antwort der Bundesregierung auf eine Kleine Anfrage der FDP- Fraktion beschäftigen sich insbesondere das „Nationale Forschungszentrum für angewandte Cybersicherheit" CRISP/ATHE-NE, aber auch die TU München oder das Fraunhofer-Institut mit diesem Thema.

Zudem haben der deutsche Auslandssender Deutsche Welle (DW), das Fraunhofer-Institut für Digitale Medientechnologie (IDMT) und das Athens Technology Center (ATC) das gemeinsame Forschungsprojekt „Digger" gestartet. Ziel des Projekts ist es, die webbasierte Verifikationsplattform „Truly Media" von DW und ATC u. a. um die Audioforensik-Technologien des Fraunhofer IDMT zu erweitern und Journalisten auf diese Weise zu helfen.[17] Hieraus lassen sich aber weder eine konkrete Strategie noch eine Investitionsabsicht der Bundesregierung ableiten.

4.2 Selbstregulierungsversuche der Social-Media-Plattformen

Während die Big-Tech-Unternehmen mit Daten und finanziellen Mitteln zu einer technologischen Lösung für die Problematik beitragen wollen, werden mehr und mehr Stimmen laut, die auch weitergehende Schritte von Facebook und Co. fordern, gerade da ihre Plattformen zur Verbreitung von Desinformation beitragen. Vor diesem Hintergrund haben sich Twitter und Facebook Ende 2019 bzw. Anfang 2020 zu ihren Plänen für den Umgang mit Deepfakes geäußert. Twitter bat im November 2019 seine Nutzerinnen und Nutzer um Feedback zu einem „Regelvorschlag für synthetische und manipulierte Medien".

Is the media significantly and deceptively altered or fabricated?	Is the media shared in a deceptive manner?	Is the content likely to impact public safety or cause serious harm?	
✓	✗	✗	Content **may** be labeled
✓	✗	✓	Content is **likely** to be labeled, or **may** be removed.
✓	✓	✗	Content is **likely** to be labeled.
✓	✓	✓	Content is **very likely** to be removed.

Twitter: Umgang mit synthetischen und manipulierten Medien: https://blog.twitter.com/en_us/topics/ company/2020/new-approach-to-synthetic-and- manipulated-media.html

twitter-approach (c) Twitter

Anfang Februar 2020 wurden die entsprechenden Regeln angekündigt:

Jedes Foto, Audio oder Video, das „erheblich verändert oder gefälscht“ wurde, um Menschen irrezuführen, wird dann entfernt, wenn Twitter der Ansicht ist, dass es ernsthaften Schaden anrichten kann – zum Beispiel indem es die physische Sicherheit von Menschen gefährdet oder „weitverbreitete Bürgerunruhen“ verursacht. Sollte dies nicht der Fall sein, können die Tweets dennoch als manipulierte Medien gekennzeichnet, Warnungen beim Versuch, die Inhalte zu teilen, ausgesprochen und die Inhalte in den Feeds der Nutzerinnen und Nutzer depriorisiert werden. Die Änderungen sind am 5. März 2020 in Kraft getreten.[18]

> *„Ob und an welchen technischen Lösungen zur Bekämpfung von Deepfakes in Deutschland und Europa gearbeitet wird, ist wenig bekannt.“*

Facebook geht einen Schritt weiter. Am 6. Januar 2020 hat Monika Bickert, Facebooks Vice President Global Policy Management, in einem Blogbeitrag angekündigt, dass künftig Deepfakes, die bestimmten Kriterien entsprechen, von der Plattform gelöscht werden sollen.[19]) Gelöscht werden sollen demnach Inhalte, die mithilfe von KI auf eine Weise bearbeitet oder synthetisiert wurden, die sie für eine Durchschnittsperson authentisch erscheinen lassen würden. Von dieser Richtlinie sollen allerdings Inhalte ausgenommen werden, bei denen es sich um Satire handelt, was einen signifikanten Interpretationsspielraum offenlässt. Interessanterweise gilt diese Richtlinie nicht für Cheap Fakes, sondern explizit nur für die KI-generierten Inhalte. Dementsprechend ist das bereits erwähnte gefälschte Video von Nancy Pelosi weiterhin auf Facebook verfügbar.[20]

Facebook räumte zwar ein, dass seine Faktenprüfer das Video als falsch eingestuft haben, lehnte es aber ab, es zu löschen, da das Unternehmen „keine Richtlinie habe, die vorschreibt, dass die Informationen, die auf Facebook gepostet werden, wahr sein müssen“.[21]

Dieser Ansatz entspricht auch Facebooks Verständnis von freier Meinungsäußerung und geht über das Thema Deepfake noch hinaus. Im Kontext der Debatte rund um politische Werbung schrieb Rob Leathern, Direktor des Produktmanagements bei Facebook, im Januar 2020 in einem Blog-Post, dass solche Entscheidungen nicht von Privatunternehmen getroffen werden sollen, „weshalb wir für eine Regulierung plädieren, die für die gesamte Branche gelten würde. In Ermangelung einer Regulierung bleibt es Facebook und anderen Unternehmen überlassen, ihre eigene Politik zu gestalten.“

Sicherlich lässt sich darüber diskutieren, ob Facebooks Auslegung der Meinungsfreiheit unter ethischen Gesichtspunkten richtig ist. Die Aussage von Rob Leathern macht aber auf ein wichtiges Thema aufmerksam – die fehlende oder zumindest lückenhafte Regulierung.

4.3 Regulierungsversuche durch Gesetzgeber

In Deutschland finden auf Deepfakes „generell- abstrakte Regelungen“ Anwendung, so die Antwort der Bundesregierung auf die bereits erwähnte kleine Anfrage der FDP-Fraktion. „Spezifische Regelungen auf Bundesebene, die ausschließlich Deep-Fake- Anwendungen erfassen oder für diese geschaffen wurden, existieren nicht. Die Bundesregierung überprüft den Rechtsrahmen auf Bundesebene fortlaufend daraufhin, ob aufgrund von technologischen oder gesellschaftlichen Herausforderungen ein Anpassungsbedarf besteht.“

Dies bedeutet, dass einige Teilaspekte der Deepfake-Problematik, zum Beispiel die Rachepornografie, vermeintlich durch existierende Gesetze implizit abgedeckt sind, jedoch ein expliziter Umgang mit manipulierten Inhalten fehlt. Dies betrifft nicht nur das spezielle Thema „Deepfakes“, sondern das gesamte Spektrum der Desinformation im digitalen Raum. Wie der Autor der Studie „Regulatorische Reaktionen auf Desinformation“[22] der Stiftung Neue Verantwortung aufzeigt, „sind bisherige Regulierungsversuche und politische Lösungsansätze [in Deutschland und Europa] kaum geeignet, um Desinformation einzudämmen“. Eine detaillierte Analyse des Status der Deepfake-Regulierungen in den USA zeigt

die Studie der Kanzlei WilmerHale „Deepfake Legislation: A Nationwide Survey“[23]. In den USA wurden explizite Deepfake-Gesetze bereits in das Strafrecht aufgenommen – zum Beispiel in Virginia, das nicht einvernehmliche Deepfake-Pornografie unter Strafe stellt, oder in Texas, wo Deepfakes, die Wählerinnen und Wähler beeinflussen sollen, unter Strafe gestellt werden. Ähnliche Gesetze wurden im September 2019 auch in Kalifornien verabschiedet.

Die wahrscheinlich weitestgehende Regulierung von Deepfakes hat Ende 2019 der chinesische Gesetzgeber vorgenommen. Die chinesischen Gesetze verlangen, dass Anbieter sowie Nutzer von Audioinformationsdiensten und online Videonachrichten alle Inhalte, die mithilfe neuer Technologien wie zum Beispiel Künstlicher Intelligenz erstellt oder verändert wurden, klar kennzeichnen. Während es durchaus überlegenswert ist, ob eine ähnliche Regulierung auch von anderen Ländern übernommen werden sollte, hinterlässt sie im Falle Chinas doch einen üblen Nachgeschmack: Die chinesische Regierung ihrerseits geht mit technologiegestützter Desinformation zum Beispiel gegen die Protestierenden in Hongkong vor, und es ist davon auszugehen, dass diese neue Regulierung als Vorwand für weitere Zensuranstrengungen genutzt werden wird.

Sicherlich ist eine wirksame Regulierung neuer technologischer Phänomene nicht ganz einfach. Auch in der Vergangenheit hat man sich hiermit immer wieder schwergetan. Das Fahren eines Autos im England des 19. Jahrhunderts zum Beispiel erforderte nach dem Locomotive Act von 1865 eine zweite Person, die dem Fahrzeug zu Fuß vorausging und eine rote Flagge schwenkte.[24]) Trotzdem gibt es Maßnahmen, die die Gesetzgeber jetzt schon ergreifen können, um dem Phänomen Deepfake entgegenzuwirken. Da es sich aktuell bei 96 Prozent der Deepfakes um nicht einvernehmliche Pornografie handelt, wäre es ein guter Anfang, diese, wie in Virginia oder Kalifornien, explizit unter Strafe zu stellen. In die gleiche Richtung sollte die Regulierung bezüglich Verleumdung, Betrug und Persönlichkeitsrechten gehen. Des Weiteren sollten die Gesetzgeber klare Richtlinien für den einheitlichen Umgang der digitalen Plattformen mit Deepfakes im Speziellen und mit der Desinformation im Allgemeinen schaffen.

Diese Maßnahmen können von der Kennzeichnung über die Limitierung der Verbreitung (Ausschluss aus Empfehlungsalgorithmen) bis zur Löschung von Deepfakes reichen. Zudem sollte die Förderung von Medienkompetenz für alle Bürgerinnen und Bürger, unabhängig vom Alter, eine Priorität darstellen. Eine angemessene Aufklärung darüber, wie Deepfakes entstehen und verbreitet werden, sollte die Bürgerinnen und Bürger in die Lage versetzen, Desinformation als solche zu erkennen und sich nicht davon in die Irre führen zu lassen.

4.4 Individuelle Verantwortung: kritisches Denken und Medienkompetenz

Kritisches Denken und Medienkompetenz sind die Grundlage für den differenzierten Umgang mit Desinformation. Sicherlich ist es unmöglich und wohl auch nicht wünschenswert, von jeder einzelnen Person zu verlangen, alles Gesehene grundsätzlich infrage zu stellen.

Und doch ist man heute mehr denn je gut beraten, das im Internet Konsumierte mit Vorsicht zu genießen. Das Einfachste, was jede und jeder unternehmen kann, wenn ein Bild, ein Video oder auch ein Text seltsam zu sein scheint, ist eine Google-Recherche: Viele der gefälschten Inhalte werden auf diese Weise schnell enthüllt, die Details der Fälschung sind genauso schnell im Umlauf.

Besonders wichtig ist dieser Schritt, wenn man gerade die Absicht hat, diesen Inhalt zu teilen, als „gefällt mir“ zu markieren oder zu kommentieren. Außerdem können wir bei Videos verstärkt darauf achten, ob das Blinzeln, der Gesichtsausdruck oder die Sprache unnatürlich wirken, ob Teile des Bildes verschwommen sind oder ob Objekte fehl am Platz erscheinen, auch wenn diese Merkmale mit dem Fortschritt der Deepfake-Technologie zunehmend verschwinden werden.

Künftig ist es denkbar, dass Browser-Add-ons, ähnlich wie die Werbeblocker, entstehen, die automatisiert manipulierte Inhalte identifizieren und die Nutzerinnen und Nutzer darauf aufmerksam machen. Diese Schritte setzen aber voraus, dass uns die Möglichkeiten der Manipulation überhaupt bewusst sind. Um dieses Bewusstsein bei seinen Bürgerinnen und Bürgern zu erreichen, setzt Finnland, das Land, das den ersten Platz in einer Studie zur Messung der Widerstandsfähigkeit[25] gegenüber Desinformation belegt, auf Bildungsangebote für die gesamte Bevölkerung – vom Kindergarten bis zum Rentenalter.

5. WIE GEHT ES WEITER?

Wie stark die konkreten Auswirkungen von Deepfakes auf Politik und Gesellschaft sein werden, ist noch nicht im Detail absehbar. Dies ist jedoch aber auch kein Grund für Untätigkeit. Wie bereits mehrmals betont wurde, sind weder die gefälschten Videos noch Desinformation als solche ein neues Phänomen – was neu ist, ist die zunehmende Einfachheit, sie zu kreieren, ihre steigende Qualität und ihre Verbreitungsmöglichkeiten.

Ein guter Lackmustest werden sicherlich die im Herbst 2020 stattfindenden Präsidentschaftswahlen in den USA sein. Trotzdem sollte die Empfehlung nicht lauten, einfach „auszuharren". Forscherinnen und Forscher, Technologieunternehmen, Journalistinnen und Journalisten, Regierungen und die Nutzerinnen und Nutzer selbst sollten keine Anstrengung scheuen, die negativen Auswirkungen der Fälschungen zu neutralisieren.

Im ersten Schritt bedarf es einer expliziten Regelung und einer konsequenten Bekämpfung der Deepfake-Pornografie, da diese bereits weit verbreitet ist und den betroffenen Frauen erheblichen Schaden zufügt.

Des Weiteren ist eine einheitliche gesetzliche Regelung des Umgangs mit manipulierten Inhalten in den Medien und auf Social-Media-Plattformen notwendig. Es sollte nicht in der Verantwortung von Facebook, Twitter, YouTube und Co. stehen, darüber zu entscheiden, bei welchen Inhalten es sich um eine freie Meinungsäußerung handelt und bei welchen die Grenzen der Meinungsfreiheit übertreten werden.

Dies ist eine Aufgabe des Gesetzgebers und des Rechtsstaats. Ersterer sollte aber wiederum nicht der Versuchung nachgeben, Deepfakes grundsätzlich zu verbieten. Jenseits der Gefahren eröffnet auch diese Technologie interessante Möglichkeiten – unter anderem für Bildung, Film und Satire. Technologie an sich ist neutral – es sind Menschen, die sie zum Nutzen oder zum Schaden einer Gesellschaft verwenden.

6. REFERENZEN

1. Bei diesem Quiz erhalten die Teilnehmer allgemeine Wissenshinweise in Form von Antworten und sie müssen ihre Antworten in Form von Fragen formulieren. Zu den deutschen Adaptionen gehörten Riskant von RTL und Der Große Preis von ZDF.
2. https://arxiv.org/pdf/1808.07371.pdf
3. https://www.facebook.com/watch/?v=10155618434657151
4. https://www.cnet.com/news/spy-reportedly-used-ai-generated-photo-to-connect-with-targets-on-linkedin/
5. https://openai.com/blog/better-language-models/
6. https://openai.com/blog/gpt-2-1-5b-release/
7. The State of Deepfakes: Landscape, Threats, and Impact, Henry Ajder, Giorgio Patrini, Francesco Cavalli, and Laurence Cullen, September 2019.
8. https://www.wsj.com/articles/fraudsters-use-ai-to-mimic-ceos-voice-in-unusual-cybercrime-case-11567157402
9. https://www.technologyreview.com/2019/10/10/132667/the-biggest-threat-of-deepfakes-isnt-the-deepfakes-themselves/
10. Zitat von https://www.edge.org/conversation/kevin_kelly-the-technium/
11. https://www.malariamustdie.com/
12. https://sfi.usc.edu/dit
13. The Partnership on AI (PAI) ist eine Organisation, die Universitäten, Forscher, NGOs und Unterneh- men mit dem Ziel verbindet, die Auswirkungen der KI und Ihren Einfluss auf die Gesellschaft besser zu verstehen: . www.partnershiponai.org
14. https://ai.facebook.com/blog/deepfake-detection-challenge/
15. https://futurism.com/darpa-68-million-technology-Deepfakes
16. https://niessnerlab.org/projects/thies2016face.html
17. https://dip21.bundestag.de/dip21/btd/19/156/1915657.pdf
18. https://blog.twitter.com/en_us/topics/company/2020/new-approach-to-synthetic-and-manipulated-media.html
19. https://about.fb.com/news/2020/01/enforcing-against-manipulated-media/
20. YouTube wiederum, eine weitere Plattform, die mit ihren Empfehlungsalgorithmen zur Viralität von Falschinformationen beiträgt, hat das besagte Video zwar gelöscht, verweigert jedoch eine klare Aussage zum künftigen Umgang mit Deepfakes.

6. REFERENZEN

21. https://www.politico.com/story/2019/05/24/facebook-fake-pelosi-video-1472413
22. https://www.stiftung-nv.de/sites/default/files/regulatorische_reaktionen_auf_desinformation.pdf
23. Matthew Ferraro, WilmerHale | Deepfake Legislation: A Nationwide Survey – State and Federal Lawmakers Consider Legislation to Regulate Manipulated Media.
24. https://sites.google.com/site/motormiscellany/motoring/law-and-the-motorist/locomotive-act-1865/
25. https://osis.bg/wp-content/uploads/2019/11/MediaLiteracyIndex2019_-ENG.pdf

GLOSSAR

ARTIFICIAL GENERAL INTELLIGENCE / STARKE KI

Bei der starken KI oder AGI handelt es sich um ein Konzept, bei dem Computersysteme ein breites Spektrum unterschiedlicher Aufgaben beherrschen und dadurch ein menschenähnliches Intelligenzniveau erreichen. Aktuell existieren solche KI-Anwendungen noch nicht. Das bedeutet, dass gegenwärtig noch kein System imstande ist, gleichzeitig Krebs erkennen, Schach spielen und Auto fahren zu können, selbst wenn für alle drei Problemstellungen bereits spezialisierte Systeme existieren. Mehrere Forschungsinstitute und Unternehmen arbeiten aktuell an der starken KI, es besteht aber keine Einigkeit darüber, ob und, wenn ja, wann diese realisiert werden kann.

BIG TECH

Der Begriff „Big Tech" wird in den Medien als Sammelbegriff für die dominierenden Unternehmen der Informationstechnologiebranche verwendet. Häufig wird er abwechselnd mit dem Begriff „GAFA" oder „Big Four" für Google, Apple, Facebook und Amazon benutzt („GAFAM", wenn auch Microsoft dazugezählt wird). Für die chinesischen Big-Tech-Unternehmen wird die Abkürzung BATX für Baidu, Alibaba, Tencent und Xiaomi verwendet.

CHEAP FAKES / SHALLOWFAKES

Cheap Fakes sind – im Gegensatz zu Deepfakes – Bild-, Audio- oder Videomanipulationen, die unter Zuhilfenahme relativ simpler Technologien kreiert werden. Als Beispiele hierfür können die Geschwindigkeitsdrosselung von Audioaufnahmen oder die Darstellung von Inhalten im veränderten Kontext genannt werden.

DARPA

Die Defense Advanced Research Projects Agency (Organisation für Forschungsprojekte der Verteidigung) ist eine Behörde des US-Verteidigungsministeriums, deren Aufgabe die Forschung und Finanzierung von wegweisenden Technologien im militärischen Bereich ist. Die von der DARPA finanzierten Projekte haben bedeutende Technologien bereitgestellt, die auch im nichtmilitärischen Bereich Verwendung finden, wie insbesondere das Internet, aber auch die maschinelle Übersetzung oder selbstfahrende Fahrzeuge.

DEEPFAKE

Deepfakes (eine Wortverschmelzung von Deep Learning und Fake, englisch für Fälschung) sind das Produkt zweier KI-Algorithmen, die in einem sogenannten Generative Adversarial Network (zu Deutsch„erzeugenden gegnerischen Netzwerk"), abgekürzt GAN, zusammenarbeiten. Die GANs können am besten als eine Möglichkeit beschrieben werden, algorithmisch neue Arten von Daten aus bestehenden Datensätzen zu generieren. So könnte ein GAN beispielsweise Tausende von Aufnahmen von Donald Trump analysieren und dann ein neues Bild erstellen, das den ausgewerteten Aufnahmen ähnelt, ohne aber eine exakte Kopie einer dieser Aufnahmen zu sein. Diese Technologie kann auf unterschiedliche Arten von Inhalten – Bild, Bewegtbild, Ton und Text – angewendet werden. Die Bezeichnung Deepfake wird aber vor allem auf Audio- und Videoinhalte angewendet.

DEEP LEARNING

Deep Learning ist ein Teilbereich des maschinellen Lernens, in dem künstliche neuronale Netze eingesetzt werden, die aus großen Datenmengen lernen. Ähnlich wie Menschen aus Erfahrungen lernen, führt ein Deep- Learning-Algorithmus eine Aufgabe wiederholt aus, um das Ergebnis nach und nach zu verbessern. Wir sprechen von Deep Learning, also einem „tiefen Lernen", weil die neuronalen Netze mehrere Schichten haben, die das Lernen ermöglichen. Durch Deep Learning können Maschinen komplexe Probleme lösen, selbst wenn sie vielfältige, unstrukturierte Datensätze verwenden.

DEEP PORN

Bei DeepPorn werden die Methoden von Deep Learning verwendet, um künstliche pornografische Bilder zu generieren.

GENERATIVE ADVERSARIAL NETWORK

„Erzeugende gegnerische Netzwerke" sind algorithmische Architekturen, die zwei neuronale Netze, ein generatives und ein diskriminierendes, verwenden. Diese beiden treten gegeneinander an (das generative Netz erzeugt die Daten und das diskriminierende Netz falsifiziert diese), um neue synthetische Datensätze zu generieren. Der Prozess wird mehrfach wiederholt, um Ergebnisse zu erzielen, die echten Daten extrem ähnlich sind. Die Netzwerke können mit

GLOSSAR

unterschiedlichen Typen von Daten arbeiten und somit für die Bild- wie auch für die Text-, Audio- oder Videogenerierung eingesetzt werden.

GPT-2

GPT-2 ist ein durch das Forschungsunternehmen OpenAI entwickeltes, auf einem künstlichen neuronalen Netz basierendes Framework, das in der Lage ist, automatisiert englischsprachige Texte zu generieren. Als Datengrundlage für GPT-2 dienen ungefähr 45 Millionen Seiten Text. Im Unterschied zu herkömmlichen Textgeneratoren setzt GPT-2 die Texte nicht aus fertigen Textblöcken zusammen und ist auch nicht auf eine bestimmte Domäne festgelegt. Es kann auf Grundlage eines beliebigen Satzes oder Textabschnitts neuen Inhalt generieren.

IBM WATSON

IBM Watson ist ein auf maschinellem Lernen basierendes System, das von IBM mit dem Ziel entwickelt wurde, in natürlicher Sprache gestellte Fragen zu verstehen und beantworten zu können. Große mediale Aufmerksamkeit hat IBM Watson erlangt, als es 2011 in der Fernseh-Quizshow Jeopardy die besten menschlichen Spieler geschlagen hat. Mittlerweile positioniert sich IBM Watson als „KI für Business“ und besteht aus einer Palette von Cloud- und Datenprodukten für diverse Branchen – von der Medizin bis zur Filmproduktion.

AI WINTER

Ein KI-Winter ist eine Periode abnehmenden Interesses und zurückgehender Forschungsgelder im Bereich der Künstlichen Intelligenz. Der Begriff wurde in Analogie zu der Idee eines nuklearen Winters geprägt. Das Technologiefeld KI hat seit den 1950er Jahren mehrere Hypes erlebt, auf die Enttäuschung, Kritik und Finanzierungskürzungen folgten.

KÜNSTLICHE NEURONALE NETZE

Künstliche neuronale Netze (Artificial Neural Networks, kurz ANN) sind Computersysteme, die vage von biologischen neuronalen Netzen inspiriert sind, die sich in den Gehirnen von Menschen und Tieren befinden. ANN „lernen“ die Ausführung von Aufgaben anhand von Beispielen, ohne mit aufgabenspezifischen Regeln programmiert zu sein. Sie können zum Beispiel lernen, Bilder zu identifizieren, die Katzen enthalten, indem sie Beispielbilder analysieren, die manuell als „Katze“ oder „keine Katze“ gekennzeichnet wurden, und die Ergebnisse zur Identifizierung von Katzen in anderen Bildern verwenden.

MASCHINELLES LERNEN

Das maschinelle Lernen ist im Wesentlichen eine Methode, bei der Algorithmen verwendet werden, um Daten zu analysieren, aus ihnen zu lernen und dann eine Vorhersage zu treffen. Anstatt also die Software mit exakten Anweisungen zur Ausführung einer bestimmten Aufgabe manuell zu programmieren, wird diese mit großen Datenmengen und Algorithmen trainiert, die ihr die Fähigkeit verleihen zu lernen, wie die Aufgabe auszuführen ist.

MICROTARGETING

Microtargeting ist eine Methode des digitalen Marketings, mit der versucht wird, Werbebotschaften für die potenziellen Kunden möglichst individuell auszuspielen. Dazu werden, je nach Plattform, demografische Merkmale, Interessen, Browsing-Historien etc. der Zielpersonen berücksichtigt. Abhängig von diesen Kriterien können unterschiedliche Personen vom gleichen Absender komplett unterschiedlich adressiert werden. Ursprünglich wurde die Methode für den Einsatz in politischen Kampagnen entwickelt, mittlerweile findet sie aber auch bei kommerziellen Kampagnen Verwendung.

PHISHING

Phishing ist eine Methode der Cyber-Attacke, bei der E-Mails als Instrument eingesetzt werden. Ziel ist es, den E-Mail-Empfänger glauben zu machen, dass die Nachricht authentisch und von Relevanz für ihn ist (zum Beispiel eine Benachrichtigung seiner Bank), und ihn damit zu motivieren, auf einen Link zu klicken oder einen Anhang herunterzuladen. Auf diese Weise können die Hacker Zugriff auf sensible Informationen wie Passwörter erlangen.

GLOSSAR

REVENGE PORN
Rachepornografie bezieht sich auf das Teilen von intimen sexuellen Darstellungen im Bild oder auf Video, ohne dass die abgebildete Person dem zugestimmt hätte. Oft wollen die Ex-Partner nach einer beendeten Beziehung auf diese Weise Rache nehmen. Drei Viertel der Opfer von Rachepornografie sind Frauen.

CHWACHE KI / SPEZIALISIERTE KI
Die Algorithmen der schwachen KI sind darauf spezialisiert, sehr konkrete Aufgaben zu erfüllen, zum Beispiel Gesichter zu erkennen, Sprache zu verstehen, Schach zu spielen. Auch wenn sie darin meist deutlich besser oder effizienter als Menschen sind, können sie eben nur diese spezifischen Probleme lösen. Alle heute existierenden Anwendungen, auch komplex erscheinende wie selbstfahrende Autos oder Sprachassistenzen, gehören in die Kategorie der schwachen KI.

SOCIAL ENGINEERING
Als Social Engineering werden Maßnahmen bezeichnet, die zu einer gezielten Beeinflussung von Menschen führen, beispielsweise um Zugang zu vertraulichen Informationen zu erlangen oder die Freigabe von Finanzmitteln zu erreichen. Die Praktik ist auch unter dem Begriff „Social Hacking“ bekannt, wenn das Ziel des Social Engineering darin besteht, den Zugang zu Computersystemen der entsprechenden Person oder Organisation zu erlangen.

SUPERINTELLIGENCE
Die Superintelligenz ist ein hypothetisches Konzept, bei dem die Künstliche Intelligenz nicht nur die intelligentesten Menschen, sondern auch die kollektive Intelligenz der Menschheit übertrifft.

VISHING
Vishing (Voice Phishing) ist eine Phishing-Methode, bei der statt einer E-Mail ein Anruf als Instrument eingesetzt wird. Der Einsatz von Deepfakes für die Stimmengenerierung kann zu einer erhöhten Effektivität dieser Methode führen.

ÜBER DEN AUTOR

Seit 1995 mit den ersten AOL-CDs online, ist Alexander seit 1999 in der Entwicklung von Internet-Projekten und Startups aktiv. Von 2003-2007 arbeitete er in Führungsfunktionen bei Bertelsmann in London, Shanghai und Toronto und entwickelte internetbasierte Geschäftsmodelle in der Medienindustrie.

Er ist Autor und Koautor von Fachbüchern über künstliche Intelligenz und Chatbots, Internet und die digitale Transformation von Unternehmen und Geschäftsmodellen. Er studierte Wirtschaftswissenschaften an der Universität St. Gallen und belegte Executive Education am INSEAD und am Massachusetts Institute of Technology (MIT).

Alexander Braun
Executive Director
E: alexander.braun@capco.com
T: +49 172 980 0455
S: @almarrone

ÜBER CAPCO

Capco, ein Unternehmen der Wipro Gruppe, ist eine globale Technologie- und Managementberatung, die sich auf die Gestaltung der digitalen Transformation in der Finanzindustrie spezialisiert hat. Mit einem wachsenden Kundenportfolio, von mehr als 100 globalen Organisationen, agiert Capco an der Schnittstelle zwischen Wirtschaft und Technologie. Indem Capco zukunftsorientierte Denkweisen mit umfassender Branchenkenntnis kombiniert, liefert das Unternehmen datengestützte End-to-End-Lösungen. Darüber hinaus treibt Capco digitale Anwendungen für das Bank- und Zahlungsverkehrswesen, die Kapitalmärkte, Wealth- und Asset-Management, den Versicherungs- und den Energiesektor voran. Capcos Innovationskraft wird durch seine Innovation Labs, seine preisgekrönte Be Yourself At Work-Kultur und seine Mitarbeitervielfalt zum Leben erweckt.

Um mehr zu erfahren, besuchen Sie www.capco.com oder folgen Sie uns auf Twitter, Facebook, YouTube, LinkedIn, Instagram und Xing.

GLOBALE STANDORTE

APAC
Bangalore
Bangkok
Gurgaon
Hongkong
Kuala Lumpur
Mumbai
Pune
Singapur

EUROPA
Berlin
Bratislava
Brüssel
Düsseldorf
Edinburgh
Frankfurt
Genf
London
München
Paris
Wien
Warschau
Zürich

NORDAMERIKA
Charlotte
Chicago
Dallas
Hartford
Houston
New York
Orlando
Toronto
Tysons Corner
Washington, D.C.

SÜDAMERIKA
São Paulo

WWW.CAPCO.COM

CAPCO
a wipro company