Künstliche Realität.

Gefahren der KI

SEMINARARBEIT RESEARCH & KNOWLEDGE

Angelina Djukić, Emil Masser,
Rahel Schuiki, Lea Wallgram
FH JOANNEUM
IND21 | SS 2023

Abstract

Künstliche Intelligenz ist längst kein Hirngespinst dystopischer Science-Fiction Filme mehr. Sie hat Einzug in die verschiedensten Aspekte des Alltags gefunden, und ob es uns bewusst ist oder nicht, sie hat nicht nur Einfluss auf unser Kaufverhalten, sondern auch auf unsere Realitätswahrnehmung. Im ersten Teil dieser Arbeit werden grundsätzliche Funktionsweisen von KI erläutert, sowie einige Anwendungsbereiche vorgestellt und auf ihre Risiken untersucht. Insbesondere liegt der Fokus dabei auf Deepfakes. Der zweite Teil der Arbeit befasst sich mit der Frage, wie einfach es ist, Menschen durch die Nutzung von AI zu täuschen. Dazu wurden Personen verschiedener Altersgruppen mit Fotos sowie mit KI-generierten Bildern konfrontiert. Anhand eines Beobachtungsrasters wurde untersucht, inwiefern KI-generierte Bilder für die Teilnehmer:innen als gefälscht erkennbar waren. Außerdem wurde eine qualitative Befragung durchgeführt, welche Ängste Menschen in Bezug auf KI haben.

Artificial Intelligence has long made its way out of dystopian science fiction movies and into reality. It has found use in many different aspects of our daily lives and influences not only consumer habits but also our perception of reality, whether we notice it or not. The first part of this paper aims to explain foundational functionalities of AI and lay out some of its fields of use, as well as the risks it brings, focusing especially on the topic of deepfakes. The second part of the paper aims to analyze how easy it is to fool people using AI tools. To achieve this, people of different age groups were confronted with both real and AI generated images. Using an observation grid, it was determined how and if the participants could correctly identify AI generated content. Furthermore, a qualitative survey was carried out to learn which fears people might have in relation to AI.

Inhalt

Einleitung

In den letzten Jahren hat es im Bereich der Künstlichen Intelligenz bemerkenswerte Entwicklungen gegeben, die jedoch auch potenzielle Gefahren mit sich bringen. Aber sind wir überhaupt dazu in der Lage, diese Gefahren zu erkennen? Und wie bewusst ist es den meisten Menschen, wie viel Einfluss KI auf ihr alltägliches Leben hat? Im Rahmen dieser Semesterarbeit liegt der Fokus auf manipulierten Medieninhalten, insbesondere durch Deepfake-Verfahren. Ziel ist es, Bewusstsein für die Gefahren zu schaffen, die von solchen Verfälschungen der Realität ausgehen

Was ist KI?

Der Versuch, Künstliche Intelligenz (in weiterer Folge mit KI abgekürzt) zu definieren, fällt schwerer aus, als vielleicht vermutet. Unter KI versteht man allgemein ein System oder eine Maschine, die „intelligent“ handelt – doch was zählt als intelligent?

Einen ersten Ansatz zur Beantwortung dieser Frage stellte der Informatik-Pionier Alan Turing bereits 1950 vor. Eine Maschine solle dann als intelligent gelten, wenn ein Mensch sie im Gespräch nicht als Maschine identifizieren könne. Diese Interviewsituation ist als Turing-Test bekannt und zählt auch heute noch zu den bekanntesten Untersuchungsmethoden zur Bewertung von KI. Für viele modernere Definitionen ist die Fähigkeit entscheidend, das menschliche Lernverhalten zu imitieren. Ein System ist intelligent, wenn es Erfahrungen sammeln, daraus lernen, und seine Handlungen basierend darauf flexibel anpassen kann.[1]

Maschinelles Lernen

Einer der bedeutendsten Teilbereiche von KI ist Machine Learning oder auch maschinelles Lernen, das Systemen ermöglicht, Muster aus eingespeisten Daten zu erkennen und ihre Leistung automatisch zu verbessern, ohne speziell dafür programmiert zu sein.[2]

Das Feld Machine Learning beinhaltet zahlreiche Arten und Techniken. Eine dieser Techniken ist Deep Learning. Sie basiert auf einem dichten Geflecht künstlicher Neuronen, bestehend aus mehreren Schichten, die ähnlich wie das menschliche Gehirn funktionieren. In die Eingabeschicht werden Daten gespeist, die Ausgabeschicht liefert ein Ergebnis. Dazwischen können zahlreiche Zwischenschichten liegen – aus dieser Vielzahl an Schichten leitet sich der Name Deep Learning ab. Deep-Learning-Systeme müssen zu Beginn trainiert werden, indem Daten in die Eingabeschicht eingespeist werden, deren Ergebnis bereits festgelegt ist. Basierend auf diesen Proben lernt das System selbst, welche Merkmale zu welchem Ergebnis führen, anstatt daraufhin programmiert zu werden. Beispielsweise können im Training Millionen von Bildern eingespeist werden, die entweder eine Katze beinhalten oder nicht. Es wird nicht vorgegeben, wie eine Katze aussieht, die KI lernt selbst, das zu erkennen.[3]

Arten von KI

Grundsätzlich kann zwischen zwei Arten von KI unterschieden werden: Starke KI und schwache KI. Unter schwacher KI versteht man Systeme, die für spezifische Anwendungen entwickelt wurden. Diese Anwendung kann mehrschichtige und komplexe Probleme beinhalten, die KI ist aber nicht dazu in der Lage, Tätigkeiten zu erfüllen, die über ihren Anwendungsbereich hinaus gehen. Starke KI hingegen soll uneingeschränkt „denken“ können und in diesem Sinne genau wie ein menschlicher Verstand funktionieren. Zurzeit gibt es noch kein KI-System, das eine solche allgemeine Intelligenz erreicht hat.[4]

1 Vgl. Bartneck u.a. 2019, S.6 ff.
2 Vgl. ebda, S.10 f.
3 Vgl. Lee/Chen 2022, S.47 ff.
4 Vgl. Bartneck u.a. 2019, S.8 f.

KI in im Einsatz

Anwendungsbeispiele und deren Risiken

Künstliche Intelligenz hat längst Einzug in die verschiedensten Aspekte des Alltags gefunden. In vielen ihrer Anwendungsbereiche hält sie sich jedoch im Hintergrund, sodass nur wenigen Menschen aktiv bewusst ist, womit sie es zu tun haben. Große Nutzer etwa von Deep-Learning-Systemen sind Internet- und Finanzunternehmen. Im Finanzbereich gibt es beispielsweise Entwicklungen in Richtung automatischer Kreditbewilligungen, die durch die Menge an verfügbaren Daten und die Einsparungen in Zeit und Kosten aufgrund des geringen Personalbedarfs möglich gemacht werden. Tech-Giants wie Amazon, Facebook und Google sind inzwischen zu führenden KI-Konzernen aufgestiegen. Mit jedem Klick werden im Internet Daten gesammelt, die verwendet werden können, um Inhalte anzupassen und Umsatz zu generieren.
Die Risiken dieser Praktik liegen auf der Hand: das Ziel der KI ist dabei immer die Gewinnoptimierung, das Wohlergehen der Nutzer wird dabei fast gänzlich außen vor gelassen. Ein weiterer Punkt ist die Tendenz zu Voreingenommenheit. Oft entsteht schon beim Training eines Deep-Learning-Systems ein Bias, indem unzureichende Daten eingespeist werden. Das ist unter anderem auf bestehende Voreingenommenheit in der Gesellschaft zurückzuführen, in der die Daten erhoben wurden.[5]

Das Problem mit Deepfakes

Spätestens seit 2018 ist das Thema Deepfakes in aller Munde. Mithilfe von Deep Learning können plötzlich Videos auf täuschend echte Weise gefälscht werden. Plötzlich ist alles möglich ¬– ein Politiker geht viral mit Aussagen, die er nie getätigt hat, eine Pornodarstellerin trägt das Gesicht einer Schauspielerin, die noch nie eine Nacktszene gedreht hat. Nicht nur Bildmaterial kann verändert werden, durch bestimmte Tools lassen sich auch Stimmen nachahmen.
2020 handelte es sich laut einem Standard-Artikel bei rund 96% aller Deepfakes um pornografisches Material. Bei solchen Aktionen gehe es oft um Macht- und Racheausübung.[6]

Die Technologie, auf der Deepfakes basieren, heißt Generative Adversarial Networks (GAN), was so viel wie „generierende gegnerische Netzwerke“ heißt. Ein solches Netzwerk setzt sich aus einem Paar neuronaler Netze zusammen, einem Fälschernetz und einem Detektivnetz. Das Fälschernetz generiert Bilder, die vom Detektivnetz auf Echtheit überprüft werden. In Folge trainiert sich das Fälschernetz selbst, um das Detektivnetz zu täuschen, ebenso trainiert sich das Detektivnetz, um Fälschungen besser zu erkennen. Dieser Prozess wird so lange wiederholt, bis sich ein Gleichgewicht eingestellt hat.
Zwar existieren Deepfake-Erkennungsprogramme, diese benötigen aber große Mengen an Rechnungsleistung und sind somit sehr teuer. Ein weiteres Problem leitet sich aus der Deep-Learning-Funktion der GANs selbst ab. Das GAN-Fälschernetz kann immer wieder auf neue Detektivnetze trainiert werden. Das führt zu einem Wettstreit, der letztendlich von der Leistungsfähigkeit der Computer entschieden wird.[7]

5 Vgl. Lee/Chen 2022, S.55 ff.
6 Vgl. Sommavilla/Stajić 2020
7 Vgl. Lee/Chen 2022, S.93f.

Was ist Realität?

Für die Definition des Realitätsbegriffs gibt es zahlreiche Ansätze. Viele gehen davon aus, dass sich die eigene Realität aus einer Vielzahl von Sinneseindrücken und Wahrnehmungen ergibt und im Gehirn gebildet wird. Somit ist auch nur das Teil der eigenen Realität, das wahrgenommen wird. Platon schildert dies im berühmten Höhlengleichnis, bei der sich die wahrgenommene Realität von gefesselten Menschen in einer Höhle rein aus Schattengebilden ergibt. In einer digitalisierten Welt gilt das gleiche Prinzip einer eingeschränkten Wirklichkeit.[8]

Durch die Verbindung dieses Realitätsbegriffes mit der Tendenz zur Voreingenommenheit bei KI-Systemen kommen weitere Risikopotentiale zum Vorschein. Was ist noch real, wenn Realität so einfach gefälscht werden kann? Wie unterscheiden wir in Zukunft zwischen „realer“ Realität und künstlicher Realität? Im Zuge dieser Arbeit wurde eine Umfrage durchgeführt, um einen Einblick zu bekommen, wie Menschen sich mit diesem Thema auseinandersetzen. Außerdem wurde ein Versuch durchgeführt, um zu überprüfen, wie leicht Täuschung durch KI-generierte Bilder tatsächlich ist.

Umfrage

Wie groß ist die Gefahr der KI auf uns? Sind sich die Menschen bewusst, was für ein Risiko die KI im Zusammenhang mit Fakes und Scams für sie darstellt? Wird sie als Bedrohung gesehen und gibt es in unterschiedliche Einschätzungen der Gefahren in verschiedenen Altersgruppen?

Um diesen Fragen auf den Grund zu gehen, haben wir uns dazu entschlossen eine Befragung und einen Versuch durchzuführen. Um einen Vergleich in verschiedenen Generationen herstellen zu können, werden 2 verschiedene Altersgruppen ausgewählt. Einerseits 20–30-Jährige, die den Umgang mit Smartphones, Laptops und co. gewohnt sind und auch über die neuesten KI-Systeme wie bspw. Chat-GPT und seinen Gefahren Bescheid wissen. Andersseits die Generation 50+, die im Umgang mit technischen Geräten nicht so versiert sind und denen KI im Zusammenhang mit Fakes evtl. noch kein Begriff sind.

Wir erhoffen uns durch diesen Versuch herauszufinden, wie die Generationen gegenüber KI eingestellt sind und wie sehr sie sich von generierten „Fakes“ hinters Licht führen lassen bzw. ob und welche Unterschiede zwischen den beiden Versuchsgruppen bestehen. Weiters möchten wir herausfinden, wie bewusst sich die Menschen der Gefahren der KI sind und für welche Schadvollen Zwecke sie eingesetzt werden kann.

8 gl. Hemmerling 2011, S.13 ff.

Vorbereitung Versuch und Befragung

Zunächst ging es darum, sich auf ein Medium festzulegen. Diese Entscheidung konnte schnell getroffen werden, da sich beide Altersgruppen viel in Sozialen Medien aufhalten und dort am häufigsten Fakenews mittels bearbeiteter oder generierter Bilder aufzufinden sind. Deshalb wird für den Versuch das Medium Bild verwendet. Mithilfe des Bildgenerators DALL-E, der eine Funktion von Chat-GPT ist, werden basierend auf selbst formulierten Textbeschreibungen einer Person durch das Computerprogramm 3-4 Bilder automatisch generiert. Bei der Textbeschreibung ist es dabei wichtig möglichst genaue Merkmale einer Person zu verwenden, um ein zufriedenstellendes Endergebnis zu erreichen.

Die Testpersonen erhalten dann eine Reihe von Bildern, bei denen sie beschreiben müssen, ob ihnen etwas auffällt oder komisch wirkt. Wichtig hierbei ist, dass die Testenden Personen vorher nicht wissen um was es geht. Im realen Leben erhalten sie auch keine Vorwarnung, dass Bilder generiert sein könnten und dadurch erhalten wir ein möglichst authentisches Feedback zu den Bildern.

In den Reihen von 17 Bildern befinden sich 4, die nicht generiert wurden. Alle anderen entspringen der Künstlichen Intelligenz. Während des Versuchs wird vom Moderator oder der Moderatorin schriftlich festgehalten, was den Testenden Personen aufgefallen ist und welche Bilder als echt oder unecht befunden werden.

Folgende Bilder wurden für den Versuch mit unseren Testpersonen verwendet:

Abb. 1: Reales Bild Nr. 1
(Quelle: https://de.depositphotos.com/stock-photos/regular-guy.html)

Abb. 2: Reales Bild Nr. 2
(Quelle: https://stock.adobe.com/de/images/junge-frau-schreit-und-zieht-eine-grimasse/59019650)

Abb. 3: Reales Bild Nr. 3
(Quelle: https://www.planet-wissen.de/kultur/religion/ostern/der-osterhase-100.html)

Abb. 4: Reales Bild Nr. 4
(Quelle: https://www.krone.at/1837309)

Abb. 5: Generierte Bilder von thispersondoesnotexist.com
(Quelle: https://thispersondoesnotexist.com)



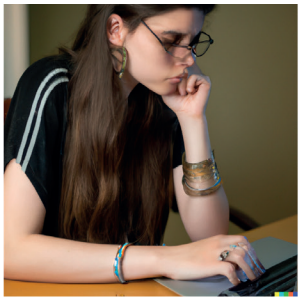








Abb. 6: Mit DALL-E generierte Bilder
(Quelle: https://labs.openai.com)

Die Bilder wurden den Testpersonen durchgemischt auf Papier ausgedruckt gezeigt.

Die Testpersonen beschreiben die Bilder und sollen Auffälligkeiten sofort mitteilen. Der Versuchsleiter oder die Versuchsleiterin notiert die Einschätzungen der Testperson auf einer Liste mit folgenden Kategorien:

1. Schöpft keinen Verdacht, das Bild wird als echt akzeptiert
2. Bringt Verwirrung/Verdacht zum Ausdruck, etwas stimmt nicht mit dem Bild
3. Erkennt das Bild als AI-generiert

Im Anschluss an den Versuch gibt es eine kurze Befragung zum Thema KI und wie die Gefahren eingeschätzt werden mit folgendem Fragebogen:

Ist dir der Begriff AI bekannt?
- o Ist mit bekannt und ich kann es genau erklären.
- o Ist mir bekannt und ich kann es grob erklären.
- o Ich kenne den Begriff, kann es aber nicht erklären.
- o Ich kenne den Begriff nicht.

Glaubst du, dass AI eine Gefahr für die Menschheit darstellen kann?
- o Ja
- o Nein
- o Ich bin nicht sicher

In welchen Bereichen siehst du Gefahren durch AI-Systeme?

(freie Antwort)

Wie besorgt bist du über die Gefahren von AI-Systemen?
- o Sehr besorgt, ich denke oft darüber nach.
- o Mäßig besorgt, ich denke manchmal darüber nach.
- o Wenig besorgt, ich denke selten darüber nach.
- o Nicht besorgt, ich denke nicht darüber nach.
- o Ich habe mich damit beschäftigt und bin überzeugt, dass AI-Systeme sicher sind.
- o Ich habe mich damit beschäftigt und bin überzeugt, dass AI-Systeme gefährlich sind.

Durchführung Versuch und Befragung

Am Test haben insgesamt 10 Personen teilgenommen, also 5 pro Altersklasse. Die Ergebnisse werden in folgende Kategorien eingeschätzt:

1. Schöpft keinen Verdacht, das Bild wird als echt akzeptiert
2. Bringt Verwirrung/Verdacht zum Ausdruck, etwas stimmt nicht mit dem Bild
3. Erkennt das Bild als AI-generiert

In den nachstehenden Tabellen werden die Ergebnisse dieses Versuchs festgehalten. Dabei wird der Durchschnittswert der Kategorien in der jeweiligen Altersklasse angeschrieben. Die Kategorien 1-3 beschreiben die oben aufgeführten Punkte der Einschätzung. Unter der Kategorie steht die Anzahl der Personen, die die Bilder in die jeweilige Kategorie eingeordnet haben. Bsp. Bild 1: 5 Personen haben das erste Bild unter Kategorie 1 eingeschätzt. Sie haben also keinen Verdacht geschöpft und das Bild wurde als echt akzeptiert.

Bild Nr.	1	2	3
Bild 1	5		
Bild 2	5		
Bild 3	5		
Bild 4	5		
Bild 5	5		
Bild 6	2	3	
Bild 7	3	2	
Bild 8	5		
Bild 9	5		
Bild 10	5		
Bild 11	5		
Bild 12	5		
Bild 13	4	1	
Bild 14	5		
Bild 15	5		
Bild 16	5		
Bild 17			

Tab. 1: Befragungsergebnisse der Kategorie 20- bis 30-Jährige

Bild Nr.	1	2	3
Bild 1	5		
Bild 2	5		
Bild 3	5		
Bild 4	5		
Bild 5	5		
Bild 6	3	2	
Bild 7	4	1	
Bild 8	5		
Bild 9	5		
Bild 10	5		
Bild 11	5		
Bild 12	5		
Bild 13	5		
Bild 14	5		
Bild 15	4	1	
Bild 16	5		
Bild 17	5		

Tab. 2: Befragungsergebnisse der Kategorie 50+Jährige

Zusammenfassung der Ergebnisse

20-30 Jährige:
-3 der Befragten machten 2 Bilder stutzig, ob sie echt sein könnten
-2 Hinterfragten kein Bild
-1 Person fand Auffälligkeiten in 3 Bildern

50+ Jährige:
-eine Person erkannte 3 Bilder als fehlerhaft
-3 Personen erkannten 2 als fehlerhaft
-für eine Person waren alle Bilder echt

Alle Altersgruppen
-6 der 10 Befragten stellten 1-2 Bilder in Frage
-Für 3 der Befragten wurden alle Bilder als echt identifiziert
-1 Person hat bei 3 Bildern Bedenken über die Echtheit

Aus den Ergebnissen ist klar zu erkennen, dass bei beiden Altersgruppen die Bilder auf ca. gleiche Weise interpretiert wurden und keine oder nur ganz selten Auffälligkeiten am Bild erkannt wurden. Obwohl bei der Auswahl der Bilder absichtlich welche benutzt wurden, die offensichtliche Fehler beinhalteten wie bspw. eine weitere Hand oder andere Pixelfehler, die beim Generieren entstanden sind. Da die Bilder fehlerhaft waren, ist es eine überraschende Erkenntnis gewesen, wie schnell die Bilder als echt festgestellt wurden. Trotz dieser offensichtlichen Fehler wurde kaum ein Bild hinterfragt. Diese Tatsache konnte noch einmal aufzeigen, wie wichtig es ist die Menschen mehr aufzuklären und ihnen bewusst zu machen, wie einfach man durch Bilder getäuscht werden kann und den kritischen Blick etwas zu schulen.

Resumé

Durch die Befragung ging hervor, dass die Gefahren durch KI von den meisten nur als sehr gering eingeschätzt werden. Gerade bei der älteren Generation sieht keiner der Befragten ein Problem in der Verwendung von KI. Sie alle gaben an, dass sie wenig bis gar nicht besorgt sind und auch nicht wirklich darüber nachdenken. Bei der jüngeren Generation deuten die Ergebnisse auf eine etwas größere Gefahr. So sind 3 der 5 Befragten sehr oder mäßig besorgt und denken öfter darüber nach. Eine Person sieht keine Gefahr und denkt auch nie darüber nach. Besonders besorgt sind beide Generationen im Bereich der Jobentwicklung und das Menschen ersetzt werden könnten. Die Angst vor Fakes, ob im Bereich der Fakenews oder Scams ist kein einziges Mal genannt worden.

Aus unserer Recherche und Umfrage konnten wir viel über den Wissensstand der Befragten Person gegenüber KI in Erfahrung bringen. So sind sich viele den Gefahren nicht bewusst und ahnen auch gar nicht, in welchen verschiedenen Bereichen die KI gefährlich sein kann. Viele der älteren Generation waren nicht darüber informiert, dass mithilfe von KI-Bilder generiert werden können die so echt aussehen, dass es von ihnen nicht identifiziert werden kann. Auch von den weiteren Bereichen wir Fakenews oder Deepfakes hatten die meisten noch nie etwas gehört. Deswegen ist es wichtig die Menschen bessere darüber zu informieren. Heutzutage kursieren überall die Geschichten von gerade älteren Personen, die auf falsche Identitäten reinfallen und Unsummen an Geld an ausländische Konten überweisen. Mit den neuen Funktionen der KI, kann diese Art der Kriminalität noch realistischer werden und mehr Opfer einfordern. Deshalb muss mehr Aufklärung in diesem Bereich betrieben werden, denn dann fallen Menschen in Zukunft nicht mehr so leicht auf generierte Medien herein.

Abbildungsverzeichnis

Literaturverzeichnis

Bartneck u.a. 2019 — Bartneck, Christoph u.a.: Ethik in KI und Robotik. München: Carl Hanser 2019

Lee/Chen 2022 — Lee, Kai-Fu/Chen, Qiufan: KI 2041. Zehn Zukunftsvisionen. Übersetzt aus dem Englischen von Thorsten Schmidt. Frankfurt am Main: Campus 2022

Sommavilla/Stajić 2020 — Sommavilla, Fabian/Stajić, Olivera: Ohne automatische Deepfake-Erkennung sind wir bald aufgeschmissen. In: Der Standard 12. Juli 2020, https://www.derstandard.at/story/2000118585546/ohne-automatische-deepfake-erkennung-sind-wir-bald-aufgeschmissen (zuletzt aufgerufen am 21.05.2023)

Hemmerling 2011 — Hemmerling, Marco: Die Erweiterung der Realität. In: Hemmerling, Marco (Hrsg.): Augmented Reality. Mensch, Raum und Virtualität. München: Wilhelm Fink 2011 (= PerceptionLab, 1)

Künstliche Realität.
Gefahren der KI