AI-Detector-Fooler

AI Noobs Unleashed

Alex Alfonso Trigo

Lou Michael Kielhorn

Lena Müller

Maximilian Olschewski

[Einleitung 3](#_Toc1908062017)

[1. Marktüberblick 3](#_Toc538978524)

[2. Funktionsweisen 4](#_Toc1948657420)

[2.1 Bildgenerierende KI 4](#_Toc850940258)

[2.2 Textgenerierende KI 5](#_Toc839430885)

[2.3 KI-Detektoren 5](#_Toc1123504920)

[2.4 Text-KI-Detektoren 5](#_Toc552358457)

[3. Aktuelle Trends 6](#_Toc1254847093)

[4. Marktherausforderungen 7](#_Toc648683393)

[5. Marktchancen 8](#_Toc1355561701)

[6. Wettbewerbsanalyse 9](#_Toc558336554)

[6.1 Bildgenerierende KI 9](#_Toc1548794608)

[6.2 Textgenerierende KI 12](#_Toc400637710)

[6.3 KI-Detektoren 14](#_Toc1315054886)

[7. Marktprognosen 16](#_Toc130648957)

[8. Fazit 16](#_Toc1602321847)

[9. Sofware Ansatz und Evaluation 17](#_Toc41083711)

[9.1 Einleitung 17](#_Toc437964776)

[9.2 Textalgorithmus 17](#_Toc1247038432)

[9.3 Bildalgorithmus 20](#_Toc1828548710)

[9.4 Browser-Extension 22](#_Toc580789658)

[9.5 Fazit 22](#_Toc1602822354)

[10. Anhang 22](#_Toc1778746932)

[Literaturverzeichnis 22](#_Toc425306879)

# Einleitung

In der heutigen schnelllebigen Ära technologischer Innovationen spielt künstliche Intelligenz (KI) eine zentrale Rolle in einer fortschreitenden digitalen Evolution. KI, als eine der markantesten Neuerungen unserer Zeit, erweitert kontinuierlich die Grenzen von maschinellem Lernen und Automatisierung. Sie setzt neue Standards für technologische Leistungsfähigkeit, von der Verbesserung alltäglicher Prozesse bis hin zur Neugestaltung komplexer Systeme in der Industrie.

KI eröffnet nicht nur innovative Wege in zahlreichen Branchen, sondern wirft auch wichtige Fragen bezüglich Ethik, Datenschutz und der Rolle des Menschen in der zukünftigen Arbeitswelt auf. In einer zunehmend datengetriebenen Welt ist KI zu einem entscheidenden Faktor geworden, der wesentliche Fortschritte in Bereichen wie Medizin, Bildung und Umweltschutz ermöglicht.

Diese Analyse liefert einen detaillierten Überblick über den Status quo der KI-Technologie, ihre vielseitigen Einsatzmöglichkeiten, die sich daraus ergebenden Marktpotenziale sowie die Herausforderungen und ethischen Fragen, die mit ihrer rasanten Entwicklung einhergehen. Ziel ist es, ein umfassendes Verständnis für die Komplexität und das transformative Potential der KI zu schaffen und einen Einblick in zukünftige Entwicklungen dieser beeindruckenden Technologie zu bieten.

# Marktüberblick

Künstliche Intelligenzen sind in ihren Anwendungsgebieten fast unbegrenzt einsetzbar. Ob nun ein Autor eine textgenerierende KI zur Hilfe nimmt, um seine Wortwahl zu verbessern oder ein Künstler eine bildgenerierende KI als Vorlage nimmt. Sie können für kreative Einzelpersonen wie diese, oder für ganze Firmen zur Analyse eingesetzt werden. Eine Firma könnte eine künstliche Intelligenz für das Marketing zur Hilfe nehmen, um vielleicht einen Slogan zu verbessern oder das Logo anzupassen. Natürlich sind künstliche Intelligenzen auch in wichtigeren Gebieten wie in Sicherheitstechnologie von Autos oder Analysen von wichtigen Daten.

Egal wie man es betrachtet, künstliche Intelligenzen spielen immer wieder eine wichtige Rolle, selbst in Gebieten, wo man es teilweise nicht mal erwartet. Für den „normalen“ Verbraucher sind Chatbots am meisten gebräuchig, wie eben ChatGPT[[1]](#footnote-2). Die Entwickler von ChatGPT ([OpenAI](https://openai.com)) haben auch die beliebte bildgenerierende KI namens DALL·E entwickelt. Diese kann ähnlich wie ein Chatbot verwendet werden, um Bilder zu generieren. OpenAI haben mittlerweile aber auch Konkurrenten im Gebiet Chatbots und Bildgenerierenden KIs. Diese sind teilweise online zugänglich, ohne etwas zu bezahlen oder in Programmen eingebettet, wie z.B. der [Copilot](https://www.microsoft.com/de-de/bing?form=MA13FV) für den Browser „Bing“.

Der Markt entwickelt sich sehr schnell, weswegen immer mehr Anbieter auf den Markt kommen werden und die Konkurrenz vorantreiben.

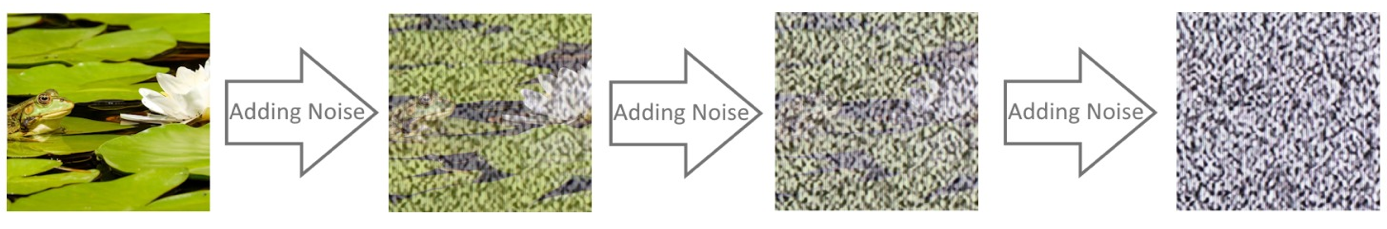
# Funktionsweisen

Eine künstliche Intelligenz besteht allgemein aus einem [neuronalen Netzwerk](https://de.wikipedia.org/wiki/Künstliches_neuronales_Netz) mit der Fähigkeit des tiefen Lernens ([Deep learning](https://de.wikipedia.org/wiki/Deep_Learning)), wodurch sie Muster in komplexen Datenmengen erkennen können[[2]](#footnote-3).

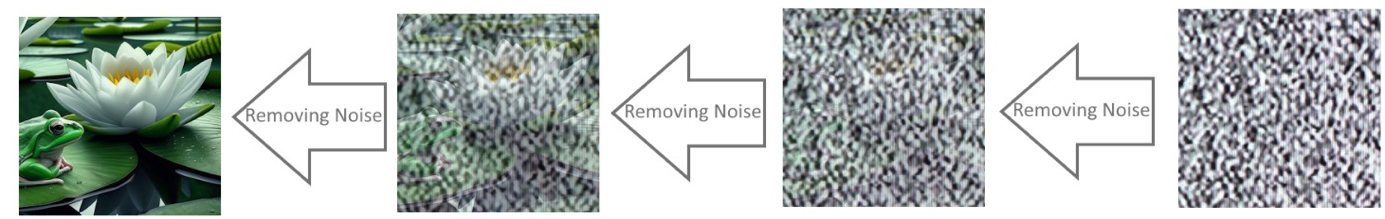
## Bildgenerierende KI

Bildgenerierende künstliche Intelligenzen benutzen solche Muster durch Embeddings. Menschen können den Zusammenhang zwischen Bild und Beschreibung leicht verstehen, jedoch muss ein neuronales Netzwerk dies in einen Vektorraum einordnen, damit es den Zusammenhang deuten kann.

Die meisten Modelle werden damit trainiert, dass bei einem Bild mit bekannter Beschreibung nach und nach ein Zufallsrauschen hinzugefügt wird. Somit lernen sie Muster aus zufälligem Rauschen zu erkennen.



Die Netzwerke werden nun darauf trainiert ein zufälliges Rauschen zu nehmen und daraus nach und nach das Zufallsrauschen zu entnehmen, damit wieder ein Bild mit der passenden Beschreibung entsteht.



(siehe Bild 1)

Die Modelle werden auf sehr großen Datensätzen trainiert mit diesem Prinzip. Mit einem umfangreichen Datensatz und einem guten Embedding sind sehr detaillierte Bilder möglich zu erstellen.

Die Bilder werden jedoch noch nicht in einer guten Auflösung erstellt. Dies wäre zu Rechen- und Zeitaufwändig. Hierfür wird ein [Upscaling](https://de.wikipedia.org/wiki/Skalierung_(Computergrafik))-Algorithmus zum Verbessern der Bildqualität benutzt, damit die Bilder auf einer für Menschen anschaulicheren Auflösung vorliegen[[3]](#footnote-4).

## Textgenerierende KI

Textgenerierende KIs, auch Large Language Models genannt (LLM)[[4]](#footnote-5) funktionieren mithilfe von maschinellem Lernen (“Machine Learning”). Dabei entwickeln sie die Fähigkeit aus Erfahrung zu lernen und automatisch Erkenntnisse und Vorhersagen aus Daten zu gewinnen[[5]](#footnote-6). Unter anderem werden hierzu künstliche neuronale Netzwerke genutzt[[6]](#footnote-7). Diese ahmen die Funktionsweise des menschlichen Gehirns nach[[7]](#footnote-8).

Sie bestehen aus mehreren Knotenschichten: der Eingabeschicht, einer oder mehrerer verdeckten Schichten und der Ausgabeschicht[[8]](#footnote-9). Sie bekommen große Datenmengen zur Verfügung gestellt und müssen, anhand von diesen, Ergebnisse zu verschiedenen Aufgaben liefern. Anhand der Rückmeldung, ob die Ergebnisse richtig sind, werden die Verbindungen zwischen den Neuronen verändert. Die Verknüpfungen, die zum richtigen Ergebnis führten, werden stärker und diejenigen, die zum Falschen führten, schwächer[[9]](#footnote-10). Diese Neuronenverbindungen ermöglichen es der KI Wahrscheinlichkeiten für Wortfolgen vorherzusagen[[10]](#footnote-11).

## KI-Detektoren

#### 2.3.1 Bild-KI-Detektoren

Bild-KI-Detektoren nutzen eine Eigenschaft, die KIs verwenden, um Bilder zu generieren. Eine KI erstellt ein Bild, indem sie ein zufälliges Rauschmuster verwendet.[[11]](#footnote-12) Dieses Rauschen wird schrittweise entfernt, bis ein erkennbares Bild entsteht. Da das Rauschmuster während der Generierung nicht vollständig entfernt wird, sondern nur für das menschliche Auge unsichtbar wird, können KI-Detektoren diese Eigenschaft nutzen. Sie werden darauf trainiert, bestimmte Rauschmuster zu erkennen und somit ein Bild als von einer KI generiert zu klassifizieren.

Einige KI-Detektoren nutzen auch banalere Merkmale. Wenn auf einer Webseite mithilfe einer KI ein Bild generiert wird, speichern die Metadaten oft die Anweisungen (Prompts) oder sogar den Namen der verwendeten KI. Diese Metadaten sind für KI-Detektoren zugänglich und ermöglichen es, ein Bild anhand dieser Informationen als von einer KI generiert zu klassifizieren. Ein Copyright-Symbol kann aber auch zum Beispiel das Gegenteil bewirken, da es häufig nur auf echten Bildern zu finden ist.

## Text-KI-Detektoren

Hinter Text-KI-Detektoren verbergen sich weitere textgenerierende KIs, welche sich den prüfenden Text anschauen und entscheiden, ob sie selbst diesen Text geschrieben haben könnten oder nicht. Falls die textgenerierende KI die Antwort Ja zurückgibt, ist der Text höchstwahrscheinlich KI-generiert.

Jedoch ist das nicht nur eine einfache Entscheidung einer textgenerierenden KI, ob der vorliegende Text KI generiert ist oder nicht, sondern solche KI hält ebenfalls nach anderen wichtigen Aspekten im Text Ausschau. Dazu gehört der Verwirrungsgrad eines Textes. Damit ist gemeint, wie sehr ein Text Verwirrung bei dem Leser auslösen kann, weil dieser unnatürlich oder ohne viel Sinn geschrieben wurde. Dazu zählen zum Beispiel Füllsätze wie "Habt ihr ebenfalls genau das Gleiche erlebt?", welche eine textgenerierende KI nicht verwenden würde. Solche Sätze oder andere unerwartete Wörter erzeugen einen hohen Verwirrungsgrad beim Leser, welcher darauf hindeutet, dass ein Mensch diesen Text geschrieben haben könnte. Textgenerierende KIs halten den Verwirrungsgrad des Textes für gewöhnlich relativ gering, da diese die offensichtlicheren Wörter wählen oder Füllwörter bzw. Füllsätze auslassen würden. Das ist eine der wichtigen Punkte, nach denen die Text-KI-Detektoren Ausschau halten.

Ebenfalls gibt es auch einen anderen Punkt, nach dem solche KIs die Texte bewerten. Dieser Punkt lautet Bustiness (zu Deutsch Burstigkeit) und das beurteilt die jeweiligen Sätze in ihrer Länge und Struktur. Bei von Menschen gemachten Texten sind die Längen und die Struktur von Satz zu Satz immer unterschiedlich, wohingegen eine textgenerierende KI Sätze konstruiert, die einer idealen Struktur folgen, wobei die Länge immer ungefähr gleich bleibt (Zum Beispiel 10-20 Wörter). Dadurch kann ein Text-KI-Detektor menschliche von KI erstellten Texten unterscheiden, da textgenerierende KIs eher zu einem Text tendieren, der eine geringere Burstiness haben als menschliche Texte.

Einen letzten Punkt überprüfen die Text-KI-Detektoren ebenfalls und diese lauten Rechtschreib- und Grammatikfehler. Menschen tendieren eher dazu, Rechtschreib- und Grammatikfehler ungewollt in ihren Texten einzubauen. Eine textgenerierende KI jedoch tut dies nicht und erzeugt nahezu jedes Mal einen fehlerfreien Text. Diesen Unterschied machen sich Text-KI-Detektoren zunutze, indem sie den Text nach Rechtschreib- bzw. Grammatikfehlern absuchen. Finden Text-KI-Detektoren einige Rechtschreib- bzw. Grammatikfehler in einem vorliegenden Text, dann ist dieser höchstwahrscheinlich von einem Menschen geschrieben.

All diese Punkte geht eine textgenerierende KI (Text-KI-Detektor) durch, bevor sie endgültig entscheidet, ob sie diesen Text selber reproduzieren könnte. Wenn die KI dennoch davon überzeugt ist, dann ist auch der vorliegende Text, der überprüft werden soll, höchstwahrscheinlich von einer KI geschrieben worden.

# Aktuelle Trends

Vom 09. Bis zum 12. Januar 2024 fand in Las Vegas die Technikmesse CES statt. Es ist die weltgrößte Messe für Verbraucher-Technologie. Im Mittelpunkt der Messe stand in diesem Jahr die künstliche Intelligenz[[12]](#footnote-13).

VW stellte seinen neuen Sprachassistenten IDA vor. Dieser basiert auf Cerence Chat Pro, ein Programm, welches verschiedene Quellen zur Beantwortung von Fragen verwendet, darunter nun auch ChatGPT[[13]](#footnote-14). Dadurch soll die Kommunikation natürlicher werden. IDA soll Fragen über die Bedienung des Fahrzeuges hinaus beantworten können[[14]](#footnote-15). Diese sollen erst von der Sprachassistenz-Software selbst beantwortet werden. Ist dies nicht möglich, wird auf ChatGPT zurückgegriffen. ChatGPT soll hierbei jedoch keinen Zugriff auf Fahrzeugdaten erhalten. Gestellte Fragen sollen direkt nach der Beantwortung gelöscht werden, um Datenschutz zu gewährleisten[[15]](#footnote-16).

Die Firma Swarovski präsentierte ein Fernglas als ihre jüngste Innovation. Dieses identifiziert rund 8000 Vogelarten, etwa 300 Säugetiere und eine Vielzahl von Schmetterlingen. Das Fernglas soll das reale Erlebnis einer Vogelbeobachtung um digitale Informationen ergänzen. Hierzu nutzt es die Merlin-Bird-ID-Datenbank der Cornell Lab of Ornithology. Auch das Erstellen von Fotos und Videos ist möglich. Durch ein integriertes GPS und den digitalen Kompass können Nutzer Orte markieren, an denen sie Tiere beobachten konnten[[16]](#footnote-17).

Die Firma Timkettle stellte ein Echtzeit-Übersetzungsgadget vor. Dieses besteht aus einem Hub und den dazugehörigen In-Ear-Kopfhörern mit Mikro. Dieses nimmt Gesagtes auf und schickt es an den Hub, woraufhin eine Übersetzung in der Cloud stattfindet. Jeder Teilnehmer eines Meetings muss einen solchen Interpreter besitzen, damit das Gesagte in die jeweilige Sprache des Teilnehmers überführt werden kann. Bei einem solchen Meeting sind bis zu 20 Teilnehmer und 5 verschiedene Sprachen möglich[[17]](#footnote-18). Bisher steht eine Auswahl von 40 Sprachen zur Verfügung.

# Marktherausforderungen

Mit dem ersten Entwurf zum AI-Act (KI-Verordnung) hat die europäische Kommission im April 2021 das erste Gesetzgebungsverfahren zu künstlicher Intelligenz in Gang gebracht. Wann dieses Verabschiedet und gültig sein wird, ist noch unklar[[18]](#footnote-19).

Die Verordnung sieht eine Klassifizierung der KIs nach Risiko vor: unannehmbares Risiko, hohes Risiko (Hochrisiko-KI-Systeme) und geringes oder minimales Risiko.

Anwendungen des unannehmbaren Risikos sind verboten. Hierzu zählen KIs, die Menschen manipulieren oder schaden, Anwendungen zur biometrischen Identifizierung von Menschen und solche, die negative Bewertungen aufgrund von sozialem Verhalten oder persönlicher Charakteristik ermöglichen.

Für Hochrisiko-KI-Systeme gelten klare Vorgaben: es muss ein Risikomanagement eingerichtet, dokumentiert und aufrechterhalten werden, Daten-Governance- und Datenverwaltungsverfahren müssen eingehalten werden und es muss eine Bewertung der Verfügbarkeit, Menge und Eignung der benötigten Datensätze durchgeführt werden. Weiterhin beinhalten sie die Führung technischer Dokumentationen, die Aufsichtsführung durch menschliches Personal und die Aufsichtspflicht über Vorgänge und Ereignisse. Nutzer müssen über die Merkmale, Fähigkeiten und Leistungsgrenzen der KI informiert werden.

Für KIs mit geringem oder minimalem Risiko gilt eine minimale Transparenz- und Informationspflicht über deren Einsatz.

Zu diesen zählen auch Chatbots. Es stellt sich die Frage, ob das Gefahrenpotenzial von KIs dieser Klassifizierung unterschätzt wird[[19]](#footnote-20).

Bildgenerierende KIs werden mit großen Datensätzen trainiert, um Bilder in verschiedensten Stilen zu generieren. Darunter befinden sich auch urheberrechtlich geschützte Werke von Künstlern. Diese sind davon wenig erfreut. Drei von ihnen haben sich in den USA zu einer Sammelklage gegen Stability AI, Deviant Art und Midjourney wegen Urheberrechtsverletzung zusammengeschlossen. Das unbefugte Einbringen in die eigenen Systeme stelle eine unzulässige Vervielfältigung dar[[20]](#footnote-21).

Data Mining, also die automatisierte Analyse von digitalen oder digitalisierten Werken, zur Gewinnung von Informationen über Muster, Trends und Korrelationen[[21]](#footnote-22), ist jedoch nicht verboten. Weder in den USA noch in Deutschland. Dies wird durch das Fair-Use-Doktrin[[22]](#footnote-23), bzw. Durch § 44 b) UrhG geregelt. Es erlaubt die Vervielfältigung von Werken für das Data Mining ohne Einwilligung. Die Daten sind jedoch nach Abschluss des Trainings zu löschen.[[23]](#footnote-24)

Doch auch Nutzer der bildgenerierenden KIs müssen aufpassen. Entstehende Werke müssen einen gewissen Abstand zum ursprünglichen Werk aufweisen. Ist dies nicht der Fall und der Nutzer veröffentlicht das Werk, ist dieser Täter einer Urheberrechtsverletzung[[24]](#footnote-25).

Die von der KI geschaffenen Werke selbst sind jedoch nicht geschützt, da die KI nicht als persönliche geistige Schöpfung gilt[[25]](#footnote-26).

# Marktchancen

Eine große Chance für den aktuellen Markt ist die medizinische Branche. Ein größeres, gutes Gebiet mit vielen Chancen ist die Radiologie. Wenn gewisse künstliche Intelligenzen darauf trainiert werden auf dem Röntgenbild Anomalien zu erkennen, die für das menschliche Auge nahezu unsichtbar sind, würde dies mehr als nur die Qualität der Diagnosen verbessern. Hierfür wird ein massiver Datensatz benötigt[[26]](#footnote-27). In Verbindung mit Erfahrung von guten Radiologen und einer guten Bilderkennung sind schnellere und sichere Befunde wahrscheinlicher[[27]](#footnote-28). Laut einer Studie sollen Deep-Learning Algorithmen Gehirntumore bereits mit einer Wahrscheinlichkeit von 98% erkennen[[28]](#footnote-29).

Eine weitere große, aber auch herausfordernde Chance ist die KI-unterstützte Bildung. Ob man nun schülerbasierte Lehrpläne oder Unterrichtsaufgaben mit einer KI generiert, eine große Möglichkeit für einen Ausbau im Markt ist da. Jedoch muss dies stark reguliert werden, zum einen muss die Qualität der KI gewährleistet sein. Zum anderen muss man aufgrund von Kinderschutz bei der Bildung noch sehr viel Zeit investieren, um tatsächlich KI-unterstütze Abläufe zu haben[[29]](#footnote-30).

# Wettbewerbsanalyse

## Bildgenerierende KI

Kostenlose bildgeneriende KIs sind für einen Einstieg in das Thema perfekt. Die meisten sind online direkt verfügbar, nachdem man sich bei den meisten Anbietern einen kostenfreien Account per Email anlegen konnte. Jedoch haben die nahezu alle Anbieter die gleichen Probleme. Man wartet teils eine Minute auf das generierte Bild, welches meistens von eher schlechter Qualität ist. Zudem hart man oft ein hartes Tages- oder Monatslimit an erstellbaren Bildern, was nicht nur das Testen erschwert hat, sondern auch für den normalen Gebraucher merkbar sein wird. Außerdem sind die Übersetzungsfähigkeiten kostenloser KIs oft zu hinterfragen.

Ein Bild, das Bär, Panda, Bild, Entwurf enthält.

Automatisch generierte Beschreibung

(siehe Bild 2)

Hier bei diesem Bild (generiert von ki-bild-erstellen.de) sieht man das erstellte Ergebnis aus dem Prompt „Panda auf dem Mond.“. Der Panda an sich ist zwar nicht allzu schlecht aussehend, jedoch ist das Bild am eigentlichen Thema vorbei, was auf ein schlechtes oder unzureichendes Embedding hinweist.

Nicht nur das Embedding kann Probleme bereiten, sondern auch unzureichende Trainingsdaten. Dies kann sich auf Bilder dadurch auswirken, dass ihnen eine gewisse Kreativität fehlt.

Ein Bild, das Riff, Text, Screenshot enthält.

Automatisch generierte Beschreibung

(siehe Bild 3)

Hier sieht man eine sich recht starke wiederholende Unterwasserlandschaft mit tropischen Fischen (generiert von imageflash.com). Bei einer viel expliziteren Eingabe wurde der Stil von Ernst Haeckel zwar beachtet, aber das Korallenriff unterscheidet sich kaum von den anderen Bildern. Zudem wurde die Vogelperspektive nicht beachtet. Dies kann ein Problem von Embeddings sein, aber auch von unzureichenden Trainingssetzen, da durch größere Trainingssetze auch größere Vielfalt und Qualität folgt.

Jedoch brauch man für einen großen Trainingssatz auch starke Rechenleistung auf dem Server. Damit folgen hohe Stromkosten und allgemein ein hohes, vorausgesetztes Budget. Die beiden Anbieter mit dem höchsten Budget die ich persönlich getestet habe war DALL·E von OpenAI und die generativeFill-Funktion von Photoshop. Beide konnten realistische und detailgetreue Bilder in verschiedenen Stilen generieren.

Ein Bild, das Bär, Säugetier, Text, Panda enthält.

Automatisch generierte Beschreibung

(siehe Bild 4)

GenerativeFill von Photoshop und DALL·E haben jeweils ihre Vor- und Nachteile.

DALL·E hat eine extrem überzeugende Bildqualität und eine sehr große „Detailliebe“.

Ein Bild, das Spielzeug, Animation, Cartoon enthält.

Automatisch generierte Beschreibung

(siehe Bild 5)

Hier hat DALL·E nur mit dem Prompt „Panda auf dem Mond“ auch noch eine Bambusflagge hinzuinterpretiert, welches sehr von Kreativität spricht.

Die Photoshop generativeFill Funktion ist dafür aber mehrseitig Einsetzbar. Sie kann bei bereits existierenden Bildern Objekte hinzufügen, entfernen oder ändern. Oder wie hier auch getestet wurde komplett neue Bilder von Grund auf Generieren. Von der Qualität sind sie ungefähr gleich Ansprechend. Bei DALL·E ist aber zudem noch ein Chatbot mit integriert, der dir Bildbeschreibungen, sowie Hilfestellungen zum besseren Generieren stellen kann. Dafür ist generativeFill extrem vielseitig einsetzbar und vielleicht besser für Künstler, Designer oder für Leute aus dem Marketing geeignet.

Je nach Benutzer würde ich also unterschiedliche Bildgeneratoren empfehlen. Für jemanden mit keiner Erfahrung kann man jegliche, kostenlose online ausprobieren, wie z.B. ImageFlash oder MyEdit Online. Wenn man jedoch ein wenig mehr generieren möchte, könnte man über ein OpenAI-Plus Abo nachdenken (20 US-Dollar pro Monat), bei dem man DALL·E unbegrenzt benutzen darf. Wenn man eher eine professionellere Umgebung haben möchte kann man sich die Photoshop Vollversion für 12 US-Dollar pro Monat holen, wodurch man die generativeFill-Funktion freischaltet.

## Textgenerierende KI

Um die Qualität erzeugter Texte durch KI zu evaluieren, habe ich vier Modelle des aktuellen Marktes ausgewählt und diese eingehend analysiert. Dabei lege ich besonderen Fokus auf die Benutzerfreundlichkeit, die Qualität des Textes, die mathematischen Fähigkeiten, die Kreativität der Plattformen und die Fähigkeit Texte zusammenzufassen. Die nachfolgende Analyse soll einen kleinen Einblick in die Kompetenzen von [ChatGPT](https://chat.openai.com/auth/login), [PopAi](https://www.popai.pro/), [Writesonic](https://writesonic.com/) und [Copy.ai](https://www.copy.ai/) bieten.

Zur Verwendung aller vier KIs ist die Erstellung eines Accounts erforderlich.

PopAi bietet verschiedene Abonnementoptionen. Der Nutzer hat die Möglichkeit zwischen verschiedenen Funktionen, wie dem Chatbot, „Chat with Document“ und „Create with AI“ zu wählen. Die Website ist klar strukturiert, über ein durchgängig eingeblendetes Textfeld ist eine schnelle Interaktion mit dem Chatbot möglich. Außerdem liefert sie eine gute Übersicht über die verschiedenen Möglichkeiten. Wird eine ausgewählt, eröffnen sich weitere. Dies erscheint im ersten Moment erfordernd, die Möglichkeiten sind jedoch sinnvoll sortiert und ermöglichen eine genaue Anpassung and die Wünsche des Nutzers.

Nach Anmeldung weist auch ChatGPT eine benutzerfreundliche Oberfläche mit einfachem Zugriff auf den aktuellen und vergangenen Chats und die Möglichkeit zum Upgrade auf GPT-4 auf. GPT-3.5 bietet keine weiteren Möglichkeiten zu dem Chatbot, was die Einfachheit der Website fördert.

Writesonic hingegen bietet unterschiedliche Abo-Optionen. Auf der Website finden sich Videoerklärungen zur Nutzung und eine klare Darstellung der Produkte und Preise. Nach dem Einloggen wird dem Nutzer die Bibliothek geöffnet, welche den Nutzer begrüßt, die Möglichkeiten bietet einen neuen Chat zu starten oder ein neues Dokument zu erstellen und einen Schnellzugriff mit allen Möglichkeiten und personalisierte Vorschläge zeigt.

Bei Copy.ai ist bereits nach der Verwendung von 2000 Worten ein Abo nötig. Auf der Website wird Copy.ai vorgestellt, nach dem Einloggen erscheint direkt der zuletzt verwendete, bzw. ein neuer Chat. Der Nutzer hat die Möglichkeit eine Datei hochzuladen, eine Vorschau von Prompt Möglichkeiten zu erhalten oder eine Brand Voice hinzuzufügen.

Alle Plattformen verfügen über eine klare Struktur und eine Benutzerfreundliche Oberfläche.

Allgemein generieren alle Plattformen einen gut lesbaren Text, der grammatikalisch korrekt ist und keine Rechtschreibfehler aufweist. Fachbegriffe werden, wenn nötig, verwendet, die Texte sind jedoch dadurch nicht komplizierter gestaltet.

Die einzigen Mankos sind, dass ChatGPT gelegentlich die generierten Stichpunkte falsch nummeriert und PopAi zum Teil auf Deutsch gestellte Fragen auf Englisch beantwortet und erst auf Nachfrage ins Deutsche übersetzt.

Um die mathematischen Fähigkeiten der KIs zu analysieren, sollten diese fünf Grundlegende mathematische Aufgaben lösen, darunter die Berechnung des Umfangs eines Kreises, die Addition von Brüchen, die Nennung aller Primzahlen bis 50 und die Berechnung eines Integrals. Bis auf eine, boten alle Plattformen für jede Aufgabe korrekte Lösungen, Writesonic war nicht in der Lage die Gleichung zu lösen. Manche der KIs boten detailliertere Lösungen als andere, PopAi und Copy.ai sind besonders herausgestochen.

Zur Überprüfung der Fähigkeit Texte zusammenzufassen hatten die KIs drei Versuche die wichtigsten Punkte aus einem Zeitungsartikel zu erfassen. Zuvor habe ich die Schlüsselpunkte herausgeschrieben, um die Ergebnisse der KIs mit diesen abzugleichen.

Keine der Zusammenfassungen enthielt alle Schlüsselpunkte. PopAi hat am besten abgeschnitten, ChatGPT und Copy.ai waren ungefähr gleich gut. Writesonic erwies sich als hilfreich, wenn die Zusammenfassung durch weitere Informationen ergänzt werden.

Im Allgemeinen sollte die Zusammenfassung der KI nach Erstellung überprüft und gegebenenfalls durch vom Nutzer als relevant betrachtete Punkte ergänzt werden.

Die Kreativität der KIs ist durch das Fortführen einer Kurzgeschichte getestet worden. Ihnen wurden drei Sätze vorgegeben, die sie unter der Erwähnung von drei Dingen um zehn Sätze ergänzen sollten. Dazu hatten sie drei Versuche, die jeweils miteinander verglichen worden sind. Aufgefallen ist hierbei, dass keine der vier KIs sich an die vorgegebenen zehn Sätze gehalten, sondern diese um eine Vielzahl von Sätzen überschritten hat.

Mit zweimal 13 und einmal 16 Sätzen hat sich Copy.ai noch am besten an die Vorgabe gehalten. Die Geschichten knüpfen gut an die vorgegebenen Sätze an und weisen keinerlei Ähnlichkeiten auf.

Auch die Geschichten von PopAi knüpfen gut an die drei Sätze an. Die Geschichten sind verschieden und auch die Namen der Charaktere kommen nicht mehrfach vor. Lediglich die Durchschnittlich 21 Sätze sind ein negativer Aspekt.

Weder die Geschichten von ChatGPT noch die von Writesonic knüpfen gut an die vorgegebenen Sätze an. Es werden mehrmals dieselben Namen verwendet und auch die Geschichten ähneln sich.

Copy.ai und PopAi lassen sich demnach als recht kreativ bezeichnen, ChatGPT und Writesonic jedoch eher weniger.

Zusammenfassend lässt sich festhalten, dass PopAi in den meisten Tests am besten abgeschnitten hat. Nutzer sollten jedoch je nach Verwendungszweck selbst abwägen, welche KI die meisten Vorteile bietet.

## KI-Detektoren

#### 6.3.1 Bild-KI-Detektoren

Der heutige Markt für Bild-KI-Detektoren lässt sich zu fünf Detektoren zusammenfassen, wobei einer besonders herausragt. Der Bild-KI-Detektor von Hivermoderation.com[[30]](#footnote-31) hat sich in unseren Tests und in vielen Forenbeiträgen als der beste KI-Bild-Detektor erwiesen.[[31]](#footnote-32)

Dies liegt vor allem daran, dass er nach dem aktuellen Stand der Technik kaum mehr überlistet werden kann, es sei denn, es erfolgt eine erhebliche Veränderung am Originalbild, die auch von einem Menschen erkannt werden könnte. Zudem sind Exploits, die vor lediglich 3 Monaten noch wirkungsvoll waren, heute nicht mehr wirksam. Dies ist größtenteils auf das engagierte Team von Hivermoderation.com zurückzuführen, das den Detektor kontinuierlich verbessert. Hivermoderation.com führt somit die Rangliste an, gefolgt von anderen KI-Bild-Detektoren, die mit ihren durchschnittlichen Ergebnissen nicht überzeugen können.

Ein Beispiel hierfür ist die Website Aiornot.com[[32]](#footnote-33), die keinen Score liefert, sondern lediglich anzeigt, ob ein Bild von einem Menschen oder einer KI erstellt wurde. Allerdings neigt dieser Detektor zu vielen falschen Aussagen und klassifiziert Bilder, die von Menschen erstellt wurden, fälschlicherweise als KI-generiert. Gleichzeitig kann man durch einfaches Hinzufügen von gaußschem Rauschen zu einem KI-generierten Bild den Detektor täuschen und vorgeben, dass das Bild von einem Menschen stammt.

Etwas bessere Detektoren sind IsItAi.com[[33]](#footnote-34), ein selbst erstellter Detektor auf Huggingface.co[[34]](#footnote-35) und Illuminarty.ai[[35]](#footnote-36). Alle drei Detektoren erkennen Bilder, die von einer KI generiert wurden und nicht nachträglich manipuliert wurden, ziemlich sicher als KI-generiert. Bei nachträglichen Änderungen variiert jedoch die Zuverlässigkeit von Bild zu Bild. Manchmal kann gaußsches Rauschen den Score auf 5% reduzieren, während in anderen Fällen nichts passiert. Der Hauptgrund, warum diese drei Detektoren bei uns nur auf dem zweiten Platz landen, liegt darin, dass sie überlistet werden können, was bei Hivermoderation.com nahezu unmöglich ist.

Insgesamt zeigt sich, dass der Bild-KI-Detektor von Hivermoderation.com aufgrund seiner beeindruckenden Unüberlistbarkeit an der Spitze steht. Trotz einiger solider Alternativen bleibt Hivermoderation.com durch seine zuverlässige Erkennung von KI-generierten Bildern und der robusten Resistenz gegenüber Täuschungsversuchen weiterhin das Maß der Dinge.

#### 6.3.2 Text-KI-Detektoren

Ebenfalls lässt sich der heutige Markt für Text-KI-Detektoren auf einige wenige Detektoren zusammenfassen, wobei auch hier einer besonders herausragt. Der Text-KI-Detektor von GPTZero.me hat sich in unserem Test als einer der besten Text-KI-Detektoren erwiesen.

Wir haben bei den einzelnen Tests versucht, den Text so zu verändern, dass dieser menschlicher wirkte, indem wir die angesprochenen Punkte in der Funktionsweise von Text-KI-Detektoren um ungefähr 10% verstärken bzw. abschwächen.

Dabei hat der Text AI Detektor GPTZero mit am besten abgeschnitten. Dieser konnte einen durchschnittlichen Score von 90% über alle Punkte hinweg halten und als wir alle Punkte gleichzeitig angepasst haben, ist er dennoch mit einem Score von 50% relativ zuverlässig. Der Erfolg von GPTZero lässt sich ebenfalls in einem anderen Text-KI-Detektor wiederfinden. Detecting-AI.com konnte ähnliche Ergebnisse erzielen, mit durchschnittlichen 75%, wenn die Punkte einzeln verändert werden oder mit einem stolzen Ergebnis von 48% als alle Punkte gleichzeitig angepasst wurden.

Dieser Erfolg lässt sich darauf zurückführen, dass diese beiden Text-KI-Detektoren unterschiedlich funktionieren als die restlichen Text-KI-Detektoren. Sowohl GPTZero als auch Detecting-AI analysieren den Text satzweise, wodurch jeder Satz einzeln in die Bewertung mit einfließt, wodurch eine bessere Einschätzung des gesamten Textes erzielt werden kann. Wir haben ebenfalls andere Text-KI-Detektoren getestet wie Scribbr.com oder Grammica.com, jedoch erzeugten diese schlechtere Ergebnisse, da sie schon bei den minimalsten Veränderungen sehr überzeugt waren, dass der Text von einem Menschen geschrieben wurde. Vor allem haben Rechtschreibfehler eine große Auswirkung gehabt, wodurch einige der Text-KI-Detektoren, nach dem Einfügen von vereinzelten Rechtschreibfehlern, den Text als 100% menschlich klassifiziert haben.

Zwischen den guten Text-KI-Detektoren und den schlechten Text-KI-Detektoren konnte ein anderer Text-KI-Detektor herausstechen, welcher ebenfalls mittelmäßige bis gute Ergebnisse liefern konnte. Der Text-KI-Detektor von Copyleaks konnte bei fast allen Punkten zuverlässig Texte von KI erkennen. Jedoch hatte auch dieser Text-KI-Detektor eine große Schwachstelle. Dies sind die Rechtschreibfehler, da er auch bei vereinzelten Rechtschreibfehlern davon überzeugt war, dass der vorliegende Text von einem Menschen geschrieben wurde. Daher reihen wir Copyleaks in der Mitte auf. Er verdient den zweiten Platz.

Zusammengefasst kann man sagen, dass sowohl GPTZero als auch Detecting-AI die Text-KI-Detektor-Szene aufgrund ihrer Zuverlässigkeit anführen, dicht gefolgt von Copyleaks, welcher ebenfalls zuverlässige Antworten liefern konnte. Dadurch bleibt immer noch eine kleine Auswahl übrig, falls man mit einem der Text-KI-Detektoren nicht zufrieden ist.

# Marktprognosen

Die Nachfrage nach Automatisierung wächst und treibt die Einführung von KI-Technologien voran. Die Größe des deutschen KI-Marktes hat eine jährliche Wachstumsrate von 15,15% (CAGR 2024-2030). Das Marktvolumen wird demnach bis 2030 wahrscheinlich auf 27,27Mrd. € steigen. KI wird in Zukunft also eine immer größere Rolle spielen.

Rechenleistung und Datenmengen werden immer größer, wodurch KIs besser trainiert und leistungsfähiger werden[[36]](#footnote-37). Die Auswahl an Gebieten, in denen sie eingesetzt werden könnten, wird immer größer. Sie könnte die Gesundheit älterer und kranker Menschen permanent überwachen, Verkehrssysteme verbessern oder den Straßenverkehr sicherer machen, da sie gefährliche Situationen schneller erkennen und auf diese reagieren kann. KI könnte auch Lehrmethoden verbessern und effektivere Nachhilfe geben, da sie sich an individuelle Vorlieben und Bedürfnisse anpassen kann[[37]](#footnote-38).

In der Industrie könnte es fahrerlose Transportsysteme geben. Auch Basistechnologien für autonome Fabriken sind möglich. Im Alltag könnten durch KI gesteuerte Roboter und Drohnen unsere Pakete liefern, die Autos könnten autonom fahren und unser Haushalt wird von Robotern geführt[[38]](#footnote-39).

# Fazit

Künstliche Intelligenz ist vielseitig einsetzbar. Es ist schwer voraussagbar, was die Zukunft für KIs bereithält, aber die Möglichkeiten sind groß. Ob im Bereich der Medizin, der Industrie oder sogar der Bildung, sie wird uns in Zukunft bei den alltäglichsten Dingen begleiten. Das wird allerdings noch einige Zeit dauern. Zuerst wird die Gesetzesgrundlage geklärt werden müssen. KI sollte nicht unreguliert bleiben. \*/ IRGENDWAS MIT ARBEITSPLÄTZEN VIELLEICHT NOCH \*/

Es bleibt mit Spannung abzuwarten, was die Zukunft für KI bereithält.

# Sofware Ansatz und Evaluation

## Einleitung

Das Ziel unseres Programmes soll es sein, die KI-Detektoren von Text und Bild zu überlisten. Dabei kann man sich mehrere verschiedene Ziele setzten. Durch den Wettbewerb ist zumindest ein Ziel bereits vorgegeben:

"Der Text bzw. das Bild soll so minimal vom Original wie möglich abweichen."

Dabei haben wir uns noch ein zusätzliches Ziel überlegt, damit dieses Projekt auch im Alltäglichen Leben Verwendung findet. Das zusätzliche Ziel ist, die Änderungen sollen, falls möglich, nicht mit dem bloßen Auge zu erkennen sein. Dadurch überlisten wir nicht nur Detektoren, sondern ermöglichen es auch Menschen zu überlisten, da diese nicht erkennen können, ob etwas am Text oder Bild aktiv verändert wurde.

Deshalb werden wir passende Methoden und Algorithmen, sowohl für die Textanpassung als auch für die Bildanpassung, suchen, welche beide Hauptziele, die wir uns gesetzt haben, am besten umsetzten und erfüllen können. Dabei ist unser größtes Hauptziel die KI-Detektoren zu überlisten.

## Textalgorithmus

Man hat bei der Marktanalyse der KI-Text-Detektoren gesehen, dass diese Falschaussagen treffen, sobald der Text Anzeichen von Grammatik- bzw. Rechtschreibfehlern aufweist. Daher versuchen wir diese Schwachstelle auszunutzen, da Grammatik- bzw. Rechtschreibfehler relativ subtil sind und wir unser erstes Hauptziel damit relativ gut umsetzten können. Dabei versuchen wir auch, dass zweite Hauptziel zu erreichen, indem wir die Veränderungen so minimal wie möglich halten.

Dabei werden wir folgende Konzepte ausprobieren und am Ende auswerten:

1.: Buchstaben werden durch passende Zahlen ersetzt [Leetspeak]:

Bei Leetspeak wird in einem Text ausgewählte Buchstaben durch Zahlen ersetzt, die sich relativ ähneln. Ein Beispiel dazu ist das O und die Null (O -> 0), dabei bleibt der Text für einen Menschen relativ gut lesbar und Wörter, in welchen Zahlen eingesetzt wurden, werden als Rechtschreibfehler gewertet, welches den Detektor überlisten würde. Dabei verstoßen wir aber gegen unser zweites Hauptziel, da so eine Änderung relativ offensichtlich ist.  
  
Bsp.:  
Original:  
Der braune Fuchs springt über den faulen Hund. Die Sonne scheint hell am Himmel und der Wald ruht in Stille. Plötzlich erwacht ein Eichhörnchen und huscht geschwind zwischen den Bäumen hindurch. Die Blätter rascheln leise im Wind, während die Natur ihre eigene Melodie spielt.  
  
Überarbeitet:  
Der braune Fuchs 5pringt über den f4ulen Hund. Die Sonne schein7 hell am Himmel und der W4ld ruht in 57ille. Plötzlich erwach7 ein Eichhörnchen und hu5cht geschwind zwi5chen den Bäumen hindurch. Die Blä7ter r4scheln lei5e im Wind, während die N47ur ihre eigen Mel0die 5piel7.

2.: Buchstaben durch ähnliche Symbole ändern:

Hierbei wird das Prinzip wie bei Leetspeak verwendet, jedoch ersetzten wir die Buchstaben nicht durch Zahlen, sondern wir nehmen passende Symbole. Ein Beispiel dazu ist das a und das @, dabei bleibt auch hier der Text relativ gut lesbar, lesbarer als bei Leetspeak, jedoch würde eine solche Veränderung relativ schnell wieder auffallen, wodurch wir wieder gegen unser zweites Hauptziel verstoßen würden. Jedoch würden die Wörter, in welchen die as durch @'s ausgetauscht werden, als Rechtschreibfehler erkannt und Sonderzeichen wirken sich ebenfalls stärker als Zahlen aus (Sonderzeichen sind seltener), wodurch eine Überlistung einer KI relativ denkbar wäre.

Bsp.:  
Original.:  
Der braune Fuchs springt über den faulen Hund. Die Sonne scheint hell am Himmel und der Wald ruht in Stille. Plötzlich erwacht ein Eichhörnchen und huscht geschwind zwischen den Bäumen hindurch. Die Blätter rascheln leise im Wind, während die Natur ihre eigene Melodie spielt.  
  
Überarbeitet.:  
Der braune Fuchs springt über den f@ulen Hund. Die Sonne scheint hell am Himmel und der W@ld ruht in Stille. Plötzlich erwacht ein Eichhörnchen und huscht geschwind zwischen den Bäumen hindurch. Die Blätter r@scheln leise im Wind, während die N@tur ihre eigene Melodie spielt.

3.: Leerzeichen durch geschützte Leerzeichen ersetzen:  
Hierbei werden Leerzeichen durch geschützte Leerzeichen ersetzt (" " -> " "). Dabei erfüllen wir unser zweites Hauptziel, da die Änderung von Leerzeichen zu Geschützen Leerzeichen kaum bis gar nicht mit dem bloßen Auge erkennbar ist. Jedoch wird der Text relativ schwer lesbar, da geschützte Leerzeichen keine Zeilenumbrüche erzeugen, wodurch lange Texte blockmäßig wirken und Wörter falsch getrennt werden. Jedoch erzeugt genau das Verhalten, dass Wörter falsch getrennt werden, Rechtschreib- bzw. Grammatikfehler, welches die KI relativ gut überlisten kann.

Bsp.:  
Original.:  
Der braune Fuchs springt über den faulen Hund. Die Sonne scheint hell am Himmel und der Wald ruht in Stille. Plötzlich erwacht ein Eichhörnchen und huscht geschwind zwischen den Bäumen hindurch. Die Blätter rascheln leise im Wind, während die Natur ihre eigene Melodie spielt.

Überarbeitet.:  
Der braune Fuchs springt über den faulen Hund. Die Sonne scheint hell am Himmel und der Wald ruht in Stille. Plötzlich erwacht ein Eichhörnchen und huscht geschwind zwischen den Bäumen hindurch. Die Blätter rascheln leise im Wind, während die Natur ihre eigene Melodie spielt.

4.:Einsatz von Homoglyphen:

Hierbei werden römische Buchstaben durch Buchstaben ersetzt aus einer anderen Sprache, die relativ ähnlich sind. Ein Beispiel dazu ist der Austausch von einem römischen "a" zu einem kyrillischen "а". Diese Änderung ist mit dem bloßen Auge kaum bis gar nicht erkennbar, welches die zweite Bedingung, die wir aufgestellt haben, erfüllt. Ebenfalls ist die Änderung über die gesamte Datei sehr minimal, da wir bloß einen Buchstaben pro ausgewähltes Wort austauschen, welches ebenfalls die erste Bedingung erfüllt. Der Nachteil ist, dass einige Schriftarten diesen Unterschied erkenntlich machen, indem sie die a's unterschiedlich darstellen. Jedoch gibt es einige Schriftarten, bei denen der Unterschied sehr minimal bis gar nicht zu erkennen ist. Dadurch das wir römische Zeichen durch kyrillische ersetzen werden die Wörter, welche aus römischen und kyrillischen Buchstaben bestehen, als Rechtschreibfehler gewertet, wodurch wir die KI relativ zuverlässig überlisten könnten. Damit könnte Homoglyphen die beste Wahl für eine Textbasierte KI sein.

Bsp.:  
Original.:  
Der braune Fuchs springt über den faulen Hund. Die Sonne scheint hell am Himmel und der Wald ruht in Stille. Plötzlich erwacht ein Eichhörnchen und huscht geschwind zwischen den Bäumen hindurch. Die Blätter rascheln leise im Wind, während die Natur ihre eigene Melodie spielt.

Überarbeitet.:  
Der braune Fuchs springt über den fаulen Hund. Die Sonne scheint hell am Himmel und der Wаld ruht in Stille. Plötzlich erwacht ein Eichhörnchen und huscht geschwind zwischen den Bäumen hindurch. Die Blätter rаscheln leise im Wind,   
während die Nаtur ihre eigene Melodie spielt.

In der Bewertung der einzelnen Algorithmen werden wir die genannten Text-KI-Detektoren verwenden, welche in der Wettbewerbsanalyse genannt wurden. Dabei werden wir die unterschiedlichen Algorithmen testen und auswerten.

Die Texte werden je nach Algorithmus zu einem anderen Prozentsatz verändert, damit man eine klare Auswirkung bei den Text-KI-Detektoren erkennen kann. Sowohl der Homoglyph- als auch der Symbol-Algorithmus verändert den Originaltext ungefähr jeweils um 3.5%. Der Leetspeak Algorithmus verändert den Text um 12.3% und der Leerzeichen Algorithmus verändert den Text ungefähr um 7.9%. Die Ergebnisse der einzelnen Algorithmen, wie gut sie jeweils bei den Text-KI-Detektoren abgeschnitten haben, sind in der folgenden Tabelle festgehalten:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Namen | Original | Leetspeak | Symbol | Leerzeichen | Homoglyphen |
| Grammica.com | 97.74% | 0,02% | 0,04% | 0,02% | 0.05% |
| GPTZero.me | 95% | 6% | 5% | 3% | 95% |
| Scribbr.com | 100% | 0% | 0% | 0% | 0% |
| Detecting-AI.com | 88.2% | 88,2% | 88,2% | 88,2% | 25,6% |

Wie es sich herausstellt, sind alle Algorithmen gleich gut, wenn diese auf unterschiedliche Prozente angepasst werden. Dadurch, dass wir jedoch das Hauptziel haben, die minimalste Abweichung vom Original zu erzeugen, wirken der Leetspeak Algorithmus und der Leerzeichen Algorithmus durch ihre im Vergleich hohe Abweichung vom Originaltext relativ unattraktiv für dieses Projekt. Daher schauen wir uns den Homoglyphen- und Symbol-Algorithmus genauer an.

Beide Algorithmen erzielen ungefähr die gleichen Ergebnisse, wobei im Durchschnitt der Symbol Algorithmus besser abschneidet. Wenn man jedoch in Betracht zieht, dass wir ebenfalls uns ein zweites Hauptziel gesetzt haben, dass der Algorithmus für das bloße Auge nicht zu erkennen sein soll, dann ist der Homoglyph-Algorithmus der ganz klare Gewinner.

Am Ende haben wir uns nach weiterem Testen für den Homoglyphen-Algorithmus entschieden, da er eher zu diesem Projekt passen würde. Er erzielt gute Ergebnisse und kann auch zuverlässig Text-KI-Detektoren überlisten. Zusätzlich erfüllt er beide von uns gesetzten Hauptziele, weil er Texte erzeugt, welche mit dem bloßen Auge nicht vom Original zu unterscheiden sind. Die Ergebnisse dieses Algorithmus variieren von Text zu Text, jedoch können wir ein zuversichtliches Ergebnis bei einigen Text-KI-Detektoren erzeugen.

## Bildalgorithmus

Bei der Aufgabe Bilder so zu verändern, dass ein KI-Detektor ein von KI generiertes Bild nicht als dieses ansieht, sind wir wie folgt herangegangen.

Da wir in vielen Forenbeiträgen gelesen haben, das KI-Detektoren das Rauschen, welches KI-Bild-Generatoren erstellen, erkennen, haben wir versucht dieses zu übertönen. Wie auch schon im Beispiel beim InformatiCup angegeben, haben wir hierfür ein gaußschem Rauschen über die Bilder gelegt.

Um unseren Algorithmus stätig testen zu können, haben wir einen Bot geschrieben, der automatisch das zu untersuchende Bild auf den verschiedenen Detektoren Webseiten hochlädt und die Scores auswertet. Dieses war leider nicht bei allen Detektoren möglich, da man sich bei manchen erst anmelden musste oder eine Abo abschließen sollte.

Mit dem nun funktionierenden Bot haben wir auch schnell erkannt, dass das gaußsche Rauschen genügt, um die meisten KI-Detektoren auszutricksen. Was man aber noch hinzufügen sollte, ist das wir zu diesem Zeitpunkt noch nicht den Hivemoderation.com KI-Detektor kannten.

Als wir uns die Metadaten der zu Verfügung gestellten Bilder angeschaut haben, ist uns schnell aufgefallen, dass in ihnen Prompts und andere Informationen gespeichert hat, die eine KI beim Generieren erstellt hat. Als wir die Metadaten entfernt hatten, hat sich tatsächlich bei manchen KI-Detektoren der Score verbessert.

Mit der Überzeugung, dass wir die Aufgabenstellung bereits gelöst haben, haben wir noch an weiteren Effekten und Veränderungen an den Bildern experimentiert in der Hoffnung, dass diese nicht so auffällig sind, wie das gaußsche Rauschen. Hier sind unsere Ergebnisse einmal aufgelistet:

|  |  |  |
| --- | --- | --- |
| **Ansatz** | **Grund** | **Wirkungsgrad** |
| Schwarze Pixel heller machen | Wenn Bilder in der Nacht mit einer Kamera aufgenommen werden, sind schwarze Pixel nie wirklich schwarz. | Gering -> max. 1-2% |
| Bild unscharf machen | KI-Generatoren erstellen häufig zu Beginn ein Rauschpattern mit dem sie Bilder generieren. Dieses soll durch die Unschärfe unkenntlich gemacht werden. | Durchschnittlich -> bei manchen Bildern bringt es viel, bei anderen gar nichts. |
| Bild mit einen Pinsel übermahlen, welcher eine Transparenz von 99% hat | Rauschpattern unkenntlich machen mit einer minimalen Veränderung des Bildes | Durchschnittlich -> bei manchen Bildern bringt es viel, bei anderen gar nichts. |
| Bild leicht schärfer machen, Bild stark schärfen | Bilder aus KI-Generatoren sehen häufig ein wenig unscharf aus. Diese Unschärfe soll beseitig werden. Der Nachteil am Stark schärfen ist, dass es eine große Veränderung am Originalbild darstellt. | Durchschnittlich bis effektiv |
| Rauschen, graues Rauschen, durchschnittliches Farbrauschen | KI-Generatoren erstellen häufig zu Beginn ein Rauschpattern mit dem sie Bilder generieren. Dieses soll durch die unkenntlich gemacht werden. Die anderen Rauschtypen bringen den gleichen Effekt. Nur für das menschliche Auge soll es nicht so schnell erkennbar sein, ob ein Bild bearbeitet wurde. | Sehr effektiv |
| Copyrighttext und Symbol hinzufügen. | Manche KI-Detektoren achten auf ein Copyrightzeichen und verringern den Score automatisch, wenn sie eins erkennen. | Bei einem KI-Detektor sehr effektiv, bei den anderen konnte der Effekt nicht festgestellt werden |
| Metadaten des Bildes entfernen oder spezielle hinzufügen. | Wenn Bilder von KI-Generatoren generiert werden, schreiben sie in die Metadaten häufig die Prompts des generierten Bildes und weitere Informationen. Die Metadaten werden von vielen KI-Detektoren auf genau solche Schlagwörter untersucht. | Bei speziellen Metadaten konnte kein Effekt festgestellt werden. Aber bei der Entfernung der Ursprünglichen konnte eine Scoreverbesserung festgestellt werden. |

Da nicht alle Ansätze eine wirkliche Verbesserung hervorbringen, haben wir uns letztendlich für einen Algorithmus entschieden, der folgendes macht:

* Bild sehr stark schärfen
* Weißer Pinsel Strich über das ganze Bild mit 25% Transparenz
* Graues Rauschen hinzufügen
* Bild als neues Bild speichern, um ursprüngliche Metadaten zu entfernen

Mit diesen Algorithmus konnten wir bis jetzt auf den meisten Webseiten einen guten Score erhalten. Lediglich bei zu kleinen Bildern zeigt er eine Schwäche, da diese aufgrund ihrer begrenzten Pixelinformationen beim Hochskalieren für das menschliche Auge deutlich verändert werden.

Aber trotz all den Ansätzen konnten wir einen KI-Detektor gar nicht überlisten. Es handelt sich hierbei um „Hivemoderation.com“. Basierend auf unseren bisherigen Erkenntnissen lässt sich feststellen, dass dieser KI-Detektor das Nonplusultra auf dem Markt ist. Er zeichnet sich dadurch aus, keine falschen Zuordnungen von Bildern als KI zu machen und zeigt eine bemerkenswerte Fähigkeit, selbst bei zahlreichen Modifikationen am Originalbild zu erkennen, ob ein Bild von einer KI generiert wurde. Vor etwa drei Monaten gab es noch einen Trick, diesen KI-Detektor zu überlisten, wie in einem [Reddit-Post](https://www.reddit.com/r/StableDiffusion/comments/11v4keo/well_i_broke_the_code_for_ai_detection/) näher erläutert wird. Die Strategie bestand darin, das Bild stark zu komprimieren, um das Rauschmuster für den Detektor unkenntlich zu machen. Aber durch die stetige Weiterentwicklung von Hivemoderation.com, funktioniert dieser Trick nun auch nicht mehr.

Zudem ist die Anwendung dieses Algorithmus für uns nicht geeignet, da er das verkleinerte Bild mithilfe einer anderen KI wieder vergrößert. Dies steht im Widerspruch zu unserer Anforderung, da das Vergrößern mehr als 5 Sekunden benötigt.

Zum Abschluss lässt sich festhalten, dass wir erfolgreich einen Algorithmus entwickelt haben, der für zahlreiche KI-Detektoren geeignet ist und das Originalbild nur in moderatem Maße verändert, um sicherzustellen, dass es nicht als völlig neues Bild wahrgenommen wird. Bedauerlicherweise variiert der Score jedoch von Bild zu Bild, wodurch wir keine durchweg herausragenden Ergebnisse erzielen konnten.

## Browser-Extension

Da wir zum Ende der Entwicklungsphase noch Zeit zur Verfügung hatten, haben wir uns dafür entschieden, eine Browser-Erweiterung zu programmieren. Diese Erweiterung kommuniziert über den NativeMessaging-Host mit unserem Python-Skript. Dadurch ist die Erweiterung stets mit dem aktuellen Algorithmus ausgestattet und benötigt keine regelmäßigen Updates.

Die Browser-Erweiterung bietet zwei grundlegende Funktionen:

Erstens ermöglicht sie es, einen Text schnell zu verändern, indem man ihn markiert und durch einen Rechtsklick auswählt oder ihn in die dafür vorgesehene Textbox einfügt. Dies ist besonders praktisch, da man häufig mit textgenerierenden KI's in Browsern interagiert und auf diese Weise schnell einen Text so verändern kann, dass er von einem Detektor nicht erkannt wird.

Zweitens ermöglicht die Erweiterung das Übergeben eines Bildes durch Drag & Drop oder durch die Auswahl eines Bildes über den Dateiexplorer. Das ausgewählte Bild wird genauso wie ein Text verändert, und das modifizierte Bild wird dann mit dem Pfad zurückgegeben.

Insgesamt ermöglicht unsere Browser-Erweiterung eine schnelle Anwendung unseres Algorithmus direkt im Browser, was für den Benutzer ein schöneres Erlebnis ist.

## Fazit

Zusammenfassend kann man sagen, dass wir bei den Ansätzen einige Möglichkeiten gefunden haben, welche die beiden aufgestellten Bedingungen mehr oder weniger erfüllen. Diese Methoden haben wir im späteren Verlauf ausreichend getestet, gegeneinander gewichtet und ausgewertet. Dabei sind wir zum Entschluss gekommen, dass bei dem Bildalgorithmus eine Mischung aus den einzelnen Algorithmen eine gute Wahl ist und bei dem Textalgorithmus ein Algorithmus besonders herausgestochen ist.

Sowohl der Text- als auch der Bildalgorithmus erzielen zufriedenstellende Ergebnisse und diese auch relativ zuverlässig. Natürlich weichen die Ergebnisse von Bild zu Bild bzw. von Text zu Text ab, jedoch sind diese Abweichungen nicht allzu stark. Daher sind wir insgesamt mit dem Ergebnis unserer Software zufrieden und mit der Umsetzung der Aufgabe. Ebenfalls haben wir die Software sowohl um eine GUI erweitert, um effizient mehrere Dateien zu konvertieren, als auch eine Browser-Extension, um das Programm effizient im Web zu benutzen. Weshalb wir die Software als vollen Erfolg zählen, auch wenn noch ein paar Sachen verbessert werden können bzw. nicht die besten Algorithmen gewählt worden sind.

# Anhang

Anhänge:  
[Bild 1](https://imgur.com/a/Gf3vdFY)

[Bild 2](https://imgur.com/a/2KM49d2)

[Bild 3](https://imgur.com/a/gON7VI5)

[Bild 4](https://imgur.com/a/GVumi5x)

[Bild 5](https://imgur.com/a/gON7VI5)

# Literaturverzeichnis

*AiOrNot*. (kein Datum). Von aiornot.com abgerufen

Amin ul Haq, Jian Ping Li, Shakir Khan, Mohammed Ali Alshara, Reemiah Muneer Alotaibi, & CobbinahBernard Mawuli. (2022). *DACBT: deep learning approach for classification of brain tumors using MRI data in IoT healthcare environment*. Scientific Reports.

Donath, A. (10. 01 2024). Von golem.de: https://www.golem.de/news/swarovski-ax-visio-fernglas-erkennt-tausende-von-tierarten-2401-180964.html abgerufen

gogodr. (2023). *reddit.com/StableDiffusion*. Von https://www.reddit.com/r/StableDiffusion/comments/11v4keo/well\_i\_broke\_the\_code\_for\_ai\_detection/https://www.reddit.com/r/StableDiffusion/comments/11v4keo/well\_i\_broke\_the\_code\_for\_ai\_detection/ abgerufen

*Hivemoderation.com*. (kein Datum). Von hivemoderation.com abgerufen

*Hivermoderation*. (kein Datum). Von hivermoderation.com abgerufen

*Huggingface*. (kein Datum). Von https://huggingface.co/spaces/umm-maybe/AI-image-detector abgerufen

IBM. (2023). Von ibm.com: https://www.ibm.com/de-de/topics/machine-learning abgerufen

*Illuminarty*. (kein Datum). Von Illuminarty.ai abgerufen

ink. (2023). Von eu-digitalstrategie.de: https://eu-digitalstrategie.de/ai-act/ abgerufen

*IsItAi*. (kein Datum). Von isitai.com abgerufen

KI PolygonSoftware. (2023). Von polygon-software.ch: https://polygon-software.ch/blog/ki-in-der-zukunft-wie-wird-sie-unser-leben-beeinflussen/ abgerufen

Kirst, N. (06. 02 2023). Von page-online.de: https://page-online.de/branche-karriere/data-mining-ist-grundsaetzlich-erlaubt-auch-fuer-kommerzielle-zwecke/ abgerufen

Krassnitzer, M. (2023). Von healthcare-in-europe.com: https://healthcare-in-europe.com/de/news/kuenstliche-intelligenz-hunger-auf-radiologie-daten.html abgerufen

Laskowski, N. (November 2023). Von TechTarget.com: https://www.techtarget.com/searchenterpriseai/definition/AI-Artificial-Intelligence#:~:text=In%20general%2C%20AI%20systems%20work,make%20predictions%20about%20future%20states. abgerufen

Linde, H. (30. Mai 2023). Von Golem.de: https://www.golem.de/news/kuenstliche-intelligenz-so-funktionieren-ki-bildgeneratoren-2305-174436-2.html abgerufen

Luber, D.-I. (. (05. 10 2023). Von bigdata-insider.de: https://www.bigdata-insider.de/was-ist-ein-large-language-model-a-d735d93bbc24d3c4091de8ce25aa36e8/ abgerufen

Manakas, M. (01. 02 2023). Von derstandard.de: https://www.derstandard.de/story/2000142990088/stable-diffusion-und-co-kann-man-fuer-die-nutzung-von abgerufen

Marx, M. (09. 01 2024). Von wiwo.de: https://www.wiwo.de/unternehmen/it/ces-2024-in-las-vegas-alle-infos-zur-groessten-tech-messe-der-welt/29591154.html abgerufen

Mercedes-Benz Group. (Juli 2023). Von genius-community.com: https://www.genius-community.com/allgemein/wie-funktioniert-eigentlich-kuenstliche-intelligenz-7867/ abgerufen

Miao, F., & Holmes, Wayne. (2023). *Guidance for generative AI in education and research.*

Müller, D. P. (02. 03 2023). Von haufe.de: https://www.haufe.de/recht/weitere-rechtsgebiete/kuenstliche-intelligenz-und-das-urheberrecht\_216\_588912.html abgerufen

Napitu, A. (27. November 2023). Von technopedia.com: https://www.techopedia.com/de/150-wichtige-statistiken-ueber-ki-einsatz-im-jahr-2023 abgerufen

progressive. (2023). Von progressivrecruitment.com: https://www.progressiverecruitment.com/de-de/blog/2018/11/kuenstliche-intelligenz-was-uns-bei-ki-zukuenftig-erwartet/ abgerufen

*Reddit.com*. (2023). Von https://www.reddit.com/r/StableDiffusion/comments/11v4keo/well\_i\_broke\_the\_code\_for\_ai\_detection/ abgerufen

Schreiber, M., & Köhn von Wedelstedt, M. (07. 12 2023). Von haerting.de: https://haerting.de/wissen/kuenstliche-intelligenz-ki-aktueller-stand-des-gesetzgebungsverfahrens-ai-act/ abgerufen

Schuldt, R., & General-Kuchel, D. (09. 01 2024). Von computerbild.de: https://www.computerbild.de/artikel/cb-News-Smart-Home-Timekettle-Interpreter-Hub-X1-Echtzeit-uebersetzer-ausprobiert-37836567.html abgerufen

Spiegel Netzwelt. (09. 01 2024). Von spiegel.de: https://www.spiegel.de/netzwelt/tech-messe-ces-vw-baut-chatgpt-in-seine-autos-ein-a-27516197-ed99-4e8c-aba9-7348bedcc09c abgerufen

Szymanski, P. (09. 01 2024). Von computerbild.de: https://www.computerbild.de/artikel/cb-News-Connected-Car-Volkswagen-integriert-ChatGPT-in-Sprachassistenten-37819053.html abgerufen

test. (2023).

Urban, E. (2022). Von t3n.de: https://t3n.de/news/ki-generierte-bilder-im-marketing-zwischen-urheberrecht-und-entscheidungshilfe-1509175/ abgerufen

Volkswagen. (08. 01 2024). Von volkswagen-newsroom.com: https://www.volkswagen-newsroom.com/de/pressemitteilungen/weltpremiere-auf-der-ces-volkswagen-integriert-chatgpt-in-aktuelle-modelle-18048 abgerufen

Weitemeyer, J. (August 2023). Von statista.com: https://de.statista.com/outlook/tmo/kuenstliche-intelligenz/deutschland#analystenmeinung abgerufen

Yogeshwar, R. (2019). „Die Daten sind das Programm“.

youknow. (22. 11 2017). Von youtube.com: https://www.youtube.com/watch?v=3RsmRMqX2IY abgerufen

1. (Napitu, 2023) [↑](#footnote-ref-2)
2. (Laskowski, 2023) [↑](#footnote-ref-3)
3. (Linde, 2023) [↑](#footnote-ref-4)
4. (Luber, 2023) [↑](#footnote-ref-5)
5. (Mercedes-Benz Group, 2023) [↑](#footnote-ref-6)
6. (IBM, 2023) [↑](#footnote-ref-7)
7. (Mercedes-Benz Group, 2023) [↑](#footnote-ref-8)
8. (IBM, 2023) [↑](#footnote-ref-9)
9. (youknow, 2017) [↑](#footnote-ref-10)
10. (Luber, 2023) [↑](#footnote-ref-11)
11. (Linde, 2023) [↑](#footnote-ref-12)
12. (Marx, 2024) [↑](#footnote-ref-13)
13. (Volkswagen, 2024) [↑](#footnote-ref-14)
14. (Szymanski, 2024) [↑](#footnote-ref-15)
15. (Spiegel Netzwelt, 2024) [↑](#footnote-ref-16)
16. (Donath, 2024) [↑](#footnote-ref-17)
17. (Schuldt & General-Kuchel, 2024) [↑](#footnote-ref-18)
18. (Schreiber & Köhn von Wedelstedt, 2023) [↑](#footnote-ref-19)
19. (ink, 2023). [↑](#footnote-ref-20)
20. (Manakas, 2023) [↑](#footnote-ref-21)
21. Urheberrechtsgesetz, § 44b [↑](#footnote-ref-22)
22. (Kirst, 2023) [↑](#footnote-ref-23)
23. (Müller, 2023) [↑](#footnote-ref-24)
24. (Müller, 2023) [↑](#footnote-ref-25)
25. (Manakas, 2023) [↑](#footnote-ref-26)
26. (Krassnitzer, 2023) [↑](#footnote-ref-27)
27. (Yogeshwar, 2019) [↑](#footnote-ref-28)
28. (Amin ul Haq, et al., 2022) [↑](#footnote-ref-29)
29. (Miao & Holmes, Wayne, 2023) [↑](#footnote-ref-30)
30. (Hivermoderation, kein Datum) [↑](#footnote-ref-31)
31. (gogodr, 2023) [↑](#footnote-ref-32)
32. (AiOrNot, kein Datum) [↑](#footnote-ref-33)
33. (IsItAi, kein Datum) [↑](#footnote-ref-34)
34. (Huggingface, kein Datum) [↑](#footnote-ref-35)
35. (Illuminarty, kein Datum) [↑](#footnote-ref-36)
36. (Weitemeyer, 2023) [↑](#footnote-ref-37)
37. (KI PolygonSoftware, 2023) [↑](#footnote-ref-38)
38. (progressive, 2023) [↑](#footnote-ref-39)