Hochschule für den
öffentlichen Dienst
in Bayern

Fachbereich
Archiv- und Bibliothekswesen

Der Einsatz von KI-Faktencheck-Tools in bibliothekarischen Schulungsveranstaltungen

Ein Leitfaden für Bibliotheken

Julia Altmann, Lilli Haubner, Sofie Henkel, Pauline Hinze

Hochschule für den öffentlichen Dienst in Bayern

Fachbereich Archiv- und Bibliothekswesen

Jahrgang 2023/2026

In Zusammenarbeit mit der AG Informationskompetenz

Inhaltsverzeichnis

1. Einleitung

In den letzten Jahren hat das Thema Künstliche Intelligenz – im Folgenden abgekürzt als KI – an großer Bedeutung gewonnen, insbesondere der Bereich generative KI. Dieser Begriff bezieht sich auf Systeme, welche selbstständig Inhalte wie Texte, Videos, Audios, Bilder und mehr erstellen können. Aufgrund einer Vielzahl an Anwendungsbereichen wird die Nutzung von generativer KI immer beliebter. Sie kann innerhalb weniger Sekunden Antworten zu Fragen geben, welche nicht mehr eigenständig recherchiert werden müssen. Neben einfachen Dingen, wie der Erstellung einer To-Do Liste, kann generative KI mittlerweile auch komplexere Anfragen bearbeiten.

In Zeiten der fortschrittlichen generativen KI wird es immer wichtiger zu erkennen, welche Inhalte von Menschen bzw. Künstlicher Intelligenz erstellt wurden. Um eine solche Erkennung zu erleichtern, gibt es seit Neuestem KI-Erkennungstools – im Folgenden auch "KI-Detektoren" oder "Tools" genannt –, welche dafür entwickelt wurden, Inhalte zu erkennen, die mithilfe von KI-Generatoren verfasst wurden. Dieser Leitfaden soll Bibliotheken einen Überblick bereits vorhandener KI-Erkennungstools bieten und die Nutzung dieser in den Kontext von Bibliotheksschulungen einbetten.

1.1 Historischer Abriss von KI-Generatoren

Künstliche Intelligenz ist ein Teilbereich der Informatik, dessen Ursprung in den 1950er Jahren liegt und vor allem durch die Nutzung von Algorithmen charakterisiert wird. Am einfachsten kann man sich einen Algorithmus als Rezept vorstellen. Er ist eine Abfolge von Schritten, die befolgt werden müssen, um ein Problem zu lösen. Künstliche Intelligenz hat das Ziel, mithilfe von Algorithmen Muster in Daten zu erkennen, um daraus Erkenntnisse zu gewinnen und sich menschliche Eigenschaften anzueignen. Im Laufe der Zeit entwickelten sich diese Systeme rasant weiter, sodass sie immer anpassungsfähiger und vielfältig anwendbar wurden. Die Fähigkeiten von KI gehen von Datenverarbeitung und Mustererkennung bis zur eigenständigen Generierung von Daten. Ein weiterer Bestandteil, der für KI eine große Rolle spielt, ist das Maschinelle Lernen. Hierbei werden Algorithmen mithilfe von tausenden oder sogar Millionen Datensätzen trainiert, um erfolgreich Aussagen über statistische Wahrscheinlichkeiten treffen zu können, welche auf Muster- und Datenerkennung basieren.

In dieser Welt der (generativen) KI wird es immer relevanter – gerade für Bibliotheken, die Nutzenden Kompetenzen wie Informationskompetenz vermitteln sollen – KI-generierte Inhalte zu erkennen, um sich sicher im Netz bewegen zu können.

1.2 Begriffsdefinition KI-Faktencheck-Tool

Was also ist ein KI-Detektor? KI-Detektoren setzen Maschinelles Lernen dafür ein, die Quelle eines gescannten Mediums (z.B. Text, Video, Audio, Bild) zu bestimmen und soll dabei helfen, zu erkennen, ob der Inhalt von Menschen oder von KI verfasst worden ist. Diese Tools werden mithilfe von Daten trainiert, die entweder von Menschen oder von KI generiert wurden. Somit hat das Tool eine Datengrundlage, basierend auf welcher mithilfe von Maschinellem Lernen Muster, Strukturen und Metadaten gesucht werden können, um Nutzenden am Ende ein Analyseergebnis liefern zu können.

Allerdings muss betont werden, dass KI-Detektoren nur Einschätzungen abgeben, welche nicht immer wahrheitsgetreu sind. In den letzten Jahren wurde die Funktionsfähigkeit solcher Tools stark verbessert, aber es gibt dennoch einige Faktoren, die den Wahrheitsgehalt einer Tool-Analyse beeinflussen:

- **Falsch-Positiv und Falsch-Negativ.** Beim Falsch-Positiv werden menschlich verfasste Inhalte fälschlicherweise als KI-generiert gekennzeichnet, während Falsch-Negative KI-generierten Content als menschlich verfasst erkennen.
- **Verzerrung.** KI-Detektoren haben teilweise Probleme damit, Daten als KI-generiert zu erkennen, wenn diese stark vermenschlicht oder modifiziert wurden. Je nach Trainingsdaten kann es vorkommen, dass Tools bei unterschiedlichen Schreibstilen, Textgattungen etc. auf Hindernisse bei der Auswertung stoßen.
- **Updates.** KI-Generatoren werden ständig weiterentwickelt. Das bedeutet, dass KI-Detektoren mit den Aktualisierungen mithalten müssen, um beispielsweise neue KI-Generatoren und deren Schreibstile zuordnen zu können.

2. Zielsetzung

Ziel dieses Leitfadens ist es, einen Überblick über vorhandene KI-Erkennungstools zu geben und die Nutzung derer im Kontext von bibliothekarischen Schulungen einzuordnen. Zusätzlich

zur Erkennung mithilfe von Tools soll die intellektuelle Prüfung der Inhalte gefördert werden, wozu einige typische Erkennungsmerkmale zur Unterstützung aufgelistet werden sollen.

Zur selbstständigen Evaluation von KI-Erkennungstools (v.a. im Rahmen von bibliothekarischen Schulungen) ist dem Leitfaden ein Katalog mit Kriterien beigefügt (vgl. Abbildungsverzeichnis), anhand derer sich die Qualität der Tools gut einschätzen und bewerten lässt.

Der Leitfaden kann je nach Bedarf als Grundlage verschiedener Schulungsveranstaltungen an Bibliotheken genutzt werden, um den Bedürfnissen und Gegebenheiten der jeweiligen Bibliothek gerecht zu werden.

3. Erkennung KI-Inhalte

Bevor ein Einblick in die KI-Erkennungstools geben wird, ist es noch einmal wichtig zu erklären, woran man von einer KI erstellte Inhalte erkennen kann. Denn trotz des Einsatzes der in diesem Leitfaden vorgestellten Tools ist es immer auch notwendig, dass der Mensch in der Lage ist, die Inhalte selbst zu prüfen, sodass im Bedarfsfall eine intellektuelle Korrektur vorgenommen werden kann. Zudem arbeiten die automatisierten Tools nicht immer zuverlässig.

Medium "Text". Ein von einer generativen KI erstellter Text lässt sich oft anhand bestimmter Muster erkennen. Das ist nicht immer ganz einfach, denn rein formal ist ein solcher nicht von einem Menschen gemachten Text zu unterscheiden. Allerdings macht oft der Inhalt deutlich, dass der Text von einer KI erstellt wurde. Vage und inhaltsarme Aussagen führen zu einem fehlenden Standpunkt. Diese fehlende Tiefe führt häufig dazu, dass sich Abschnitte und Inhalte wiederholen, ohne dass evtl. wichtige Details hinzugefügt werden. Des Weiteren basieren KI-Texte auf der Wahrscheinlichkeit, wie häufig in der realen Welt welche Wörter nacheinander aufgereiht auftauchen. Die KI folgt daher gängigen sprachlichen Mustern und benutzt vorhersehbare Wörter. Floskeln und Einleitungsphrasen häufen sich mehr in KI-generierten Texten, welche als Lückenfüller in solchen Texten dienen.

Hier ein kurzer Abriss an Fragen, die man einem Text stellen kann, um zu überprüfen, ob er KI-generiert ist. Wenn folgende Fragen angekreuzt werden können, ist es wahrscheinlich, dass der Text KI-generiert ist:

- ☐ Wirkt der Text "ruckartig"? Gibt es keine Mischung von langen und kurzen Sätzen?

- ☐ Werden vorhersehbare Wörter benutzt?
- ☐ Mehren sich inhaltsarme Aussagen? Oder doppeln sich Aussagen, nur um den Text in die Länge zu ziehen?
- ☐ Bleibt der Text oberflächlich? Wiederholen sich Strukturen?
- ☐ Enthält der Text keine Rechtschreibfehler, die bei einem menschengemachten Text erwartbar sind?

Medium "Bild". Generativ erstellte Bilder zu erkennen ist wichtiger denn je. Gerade extreme Strömungen in den Sozialen Netzwerken nutzen diese Art Bilder zu verbreiten als Propagandawerkzeug. Denn: "Ein Bild sagt mehr als tausend Worte". Wir schenken besonders Bildern und Videos besondere Glaubwürdigkeit, da wir glauben, unseren Augen vertrauen zu können. Daher ist es hier besonders wichtig zu wissen, woran man KI-generierte Bilder erkennt. Bei Bildern liegt das Augenmerk oftmals im Detail der Bildlogik. Bei KI-generierten Bildern häufen sich Fehler, die so in der Realität nicht vorkommen. Hier ist es besonders hilfreich, einzelne Objekte genau zu betrachten und nachzuverfolgen. Hält beispielsweise eine Person eine Schnur in der Hand, die dann aber im Nichts verläuft oder sich mit anderen Gegenständen vermischt, wäre das ein Indiz darauf, dass das vorliegende Bild KI-generiert wurde. Auch wenn im Bild Text und/oder Schrift auftaucht, welcher allerdings nicht richtig dargestellt wird, kann das ein Indiz für eine KI als Urheber sein. Denn eine KI kann oftmals keine korrekte Schrift oder Buchstabenkombination bilden. Auch sonst hat sie Probleme mit Buchstaben und Worten.[1] Die Folgen sind fehlerhafte, verschwommene und/oder unlesbare Texte, evtl. auch in Spiegelschrift.

Hier ein kurzer Abriss an Fragen, die man einem Bild stellen kann, um zu überprüfen, ob es KI-generiert ist. Wenn folgende Fragen angekreuzt werden können, ist es wahrscheinlich, dass das Bild KI-generiert ist:

- ☐ Gibt es unnatürliche Darstellungen von Körperteilen (z.B. zu wenige/zu viele Finger an Händen)?
- ☐ Scheinen Licht- und Schattenverhältnisse unnatürlich? Werden Spiegelungen und Reflexionen unrealistisch dargestellt?

[1] Stand 2025

- ☐ Sind Umrisse von Körpern unscharf?
- ☐ Wirken Gesichter zu perfekt (z.B. ungewöhnlich glatte Haut oder wenig asymmetrische Gesichtszüge)?
- ☐ Verändern zusammenhängende Formen und Objekte wie z.B. Ketten, Schnüre, Kleidungsstücke etc. Farbe, Textur oder Form?
- ☐ Sind die Texturen (Musterung von Böden, Wänden, etc.) inkonsistent?
- ☐ Ist die Perspektive unstimmig?

Medium "Audio". Eine Audiodatei zu fälschen ist einfacher als das bei Bild- oder Videodateien der Fall ist, denn schon mit einer geringen Anzahl an Trainingsdaten kann man schon gute Ergebnisse erzielen. Hinzu kommt, dass man bei Audios weniger Anhaltspunkte besitzt als bei Bildern oder Videos, was der Erkennung eine besondere Schwierigkeit beschert. Soundeffekte oder Musikaufnahmen sind in diesem Bereich hervorzuheben. Insbesondere Deepfake-Audios zu erkennen, ist jedoch von höchster Wichtigkeit, denn nicht selten wird versucht mithilfe von KI-generierten Audios Fake-News (z.B. bei Wahlen) zu verbreiten oder auch Betrüge (z.B. am Telefon) durchzuführen. Um dem zu entgehen, lohnt es sich genau hinzuhören und die eigenen Sinne zu trainieren. Falsche Aussprache kann ein Indiz für KI-generierte Inhalte sein. Zum Beispiel wird ein nur mit englischer Sprache trainiertes Modell Schwierigkeiten aufweisen, deutsche Begrifflichkeiten richtig auszusprechen und andersherum. Auch lohnt sich eine Recherche nach der Original-Stimme und ein Vergleich mit dem Deepfake, denn oft unterscheiden sich die Beiden in Stimmmonotonie, Betonung der genutzten Wörter oder auch Sprachmelodie. Auch bei Audios – wie schon bei Texten – sollte man auch immer auf Plausibilität des Inhalts überprüfen.

Hier ein kurzer Abriss an Fragen, die man einer Audiodatei stellen kann, um zu überprüfen, ob sie KI-generiert ist. Wenn folgende Fragen angekreuzt werden können, ist es wahrscheinlich, dass die Audiodatei KI-generiert ist:

- ☐ Klingt die Datei metallisch/mechanisch?
- ☐ Gibt es eine falsche Aussprache?
- ☐ Wird eher monoton gesprochen? Werden die Wörter und/oder Sätze nicht korrekt betont?

- ☐ Wird emotionslos gesprochen? Gibt es keine “natürlichen Pausen” (z.B. durch “ähm” oder ähnliche Füllwörter)?
- ☐ Gibt es unnatürliche Geräusche (im Hintergrund) oder Verzögerungen in der Artikulation?
- ☐ Beim Vergleich des Audioinhalts ist das Ergebnis nicht das Gleiche?

Medium “Video”. Bei Videos ist es wichtig, auf Details und Unstimmigkeiten im Bildmaterial und Ton zu achten. Bei sogenannten Deepfake-Videos wird entweder eine digitale Maske über das Gesicht gezogen, oder es wird die Mimik und Mundbewegung der Person im Video von einer KI lediglich so verändert, dass beides zum neuen Ton passt. In diesen Fällen sollte man besonders die Lippen- und Mundpartie beachten. Auch mit Zähnen haben KIs beim Erstellen Probleme. Ist ein Video komplett von einer KI erstellt worden, weisen diese Videos ähnliche Probleme wie Bilder auf: Objekte vermischen sich mit anderen Objekten, verschwinden oder ändern ihre Form. Außerdem hat auch hier KI einen Hang zur Perfektion. Besonders in Gesichtern von Lebewesen kommt es zu Auffälligkeiten.

Hier ein kurzer Abriss an Fragen, die man einem Video stellen kann, um zu überprüfen, ob es KI-generiert ist. Wenn folgende Fragen angekreuzt werden können, ist es wahrscheinlich, dass das Video KI-generiert ist:

- ☐ Wirken Lebewesen mechanisch, leblos? Machen sie langsame Bewegungen? Bleibt eine Person zu statisch (keine kleinen Gesten, keine Augen- bzw. Gesichtsbewegungen)?
- ☐ Passieren im Video physisch unmögliche Dinge (z.B. Laufen in die falsche Richtung)?
- ☐ Wirken Lebewesen und Gegenstände zu perfekt?
- ☐ Für Deepfakes: passt etwas nicht zusammen (z.B. Lippenbewegung zu Audiospur, Alterung der Gesichtsmerkmale)?
- ☐ Blinzelt eine Person zu häufig oder auch zu selten?
- ☐ Gibt es unrealistische Gesamteindrücke (z.B. unrealistische Gesichtsbehaarung oder Muttermale)?

Bei all diesen Medien muss beachtet werden, dass sich KI rasend schnell weiterentwickelt. Gesetzmäßigkeiten, welche heute gelten, müssen nicht zwangsläufig in einem Jahr noch ihre

Gültigkeit haben. Auch die KI-Tools, die auf Erkennung von KI-Content spezialisiert sind, entwickeln sich weiter und können als Stütze herangezogen werden.

4. Auswahl an gesichteten Tools

Im Folgenden wird – geordnet nach Medientyp – eine Auswahl an nutzbaren Tools aufgelistet. Manche der Tools können mehrere Medientypen erkennen. Das Tool Deep-Fake-o-Meter wurde für die Bereiche Audio und Video getestet und funktioniert bei beiden Bereichen auf ähnliche Weise, weshalb es nur einmal aufgelistet wurde. Hive Moderation, das für Bilder, Videos und Audios geeignet ist, wurde dagegen dreimal aufgeführt, da es je nach Medientyp verschiedene Funktionen bietet.

4.1 Text

GPTZero[2] ist ein KI-Detektor, welcher darauf spezialisiert ist, akademische Texte zu erkennen, die von KI-Modellen wie ChatGPT verfasst wurden. Das Tool wurde mit einer Vielzahl an menschen- und KI-generierten Texten trainiert und kann KI auf Satz-, Absatz- und Dokumentebene erkennen. Auf der Website sind vergangene Aktualisierungen dokumentiert, in welchen Features verzeichnet sind, die neu, verbessert oder entfernt wurden.[3] Neben dem KI-Detektor bietet GPTZero auch Tools wie Plagiatsprüfer und Grammatikprüfer an.

Isgen.ai bezeichnet sich als den „genauste[n] deutsche[n] KI-Detektor“[4], welcher KI-Inhalte in über 80 Sprachen auf Wortebene erkennen kann. Bereits für den kostenfreien Plan können 12.000 Wörter pro Monat in Form einer grundlegenden KI-Erkennung gescannt werden. Neben dem KI-Detektor bietet das Unternehmen auch Tools wie Plagiatsprüfer, Zitationsgenerator oder Grammatikprüfer an. Eine Besonderheit bei Isgen.ai ist die Möglichkeit, im KI-Detektor einen Beispieltext aus bekannten KI-Modellen wie ChatGPT, Gemini etc. für den Scan generieren lassen zu können.

Das Tool **ZeroGPT**[5] spezialisiert sich vor allem auf Texte, die von ChatGPT und GPT-4 verfasst wurden. Neben dem KI-Detektor können Texte auf Plagiat überprüft, zusammengefasst,

[2] https://gptzero.me/
[3] https://gptzero.notion.site/GPTZero-Release-Notes-Model-and-API-6f58686f6381498baef35212463b7da6
[4] https://isgen.ai/de
[5] https://www.zerogpt.com/

verbessert, paraphrasiert und übersetzt werden. Zusätzlich zur Website ist es Nutzenden möglich, das Tool als WhatsApp-Kontakt hinzuzufügen und dort beliebige Texte mithilfe des Signalworts „detect“ auf KI prüfen zu lassen. Das Tool gibt an, für jede Personengruppe nützlich zu sein. Allein in der kostenfreien Version können pro Prompt jeweils 15.000 Wörter gescannt werden.

Smodin[6] umfasst Tools, welche sich rund um die Arbeit mit KI-generierten Texten drehen, sei es ein KI-Detektor, Plagiatsprüfer oder KI-Vermenschlicher. Diese Angebote richten sich hauptsächlich an Studierende, Autor*innen und Unternehmer*innen in mehr als 180 Ländern der Welt. Smodin behauptet, andere KI-Detektoren Tools in den Bereichen Präzision, Genauigkeit und Geschwindigkeit bei Weitem zu übertreffen.[7]

QuillBot [8] wurde 2017 gegründet und hilft Nutzenden dabei, deren Schreibprojekte zu verbessern. Die Angebote sind nicht nur auf KI-generierte Inhalte fokussiert, sondern erledigen auch allgemeine Aufgaben wie Übersetzungen, Zusammenfassungen oder Verbesserungen der Grammatik. QuillBot kann in mehreren Browsern als Erweiterung eingestellt werden und somit bei jeder alltäglichen Schreibaufgabe helfen. Der KI-Detektor kann insgesamt sechs Sprachen erkennen. Im Eingabefenster des Tools können beliebig viele Wörter eingegeben und gescannt werden.

4.2 Bild

Winston AI[9] lobt sich selbst mehrfach als „[den] vertrauenswürdigste[n] AI-Detektor“. Der Detektor kann für KI-generierte Inhalte der Medientypen Text und Bild sowie zur Erkennung von Plagiat verwendet werden. Auf seiner Homepage spricht das Tool mit seinen Diensten gezielt den akademischen Sektor, Verlage sowie Schriftsteller an. Es ist möglich, bei Interesse Partner- und Botschafterprogrammen von Winston AI beizutreten.

[6] https://smodin.io/de

[7] https://smodin.io/de/Fallstudien/smodin-ai-detektor-gegen-konkurrenten

[8] https://quillbot.com/de/

[9] https://gowinston.ai/de/

Illuminarty[10] dient der Erkennung von KI-generierten Inhalten in Text und Bild. In seiner erweiterten Version zeigt es zusätzlich zum Scanergebnis, welches KI-Modell zur Generierung des erkannten Inhalts genutzt wurde. Laut eigener Website ist aktuell eine Browser Erweiterung in Arbeit, welche eine Text- oder Bildüberprüfung direkt mit einem Mausklick erlauben soll.

Is It AI? [11] verfügt aktuell über zwei verschiedene Detektoren. Während der Text-KI-Inhaltsdetektor Texte auf ihren logischen Zusammenhang prüft, analysiert der Bild-KI-Inhaltsdetektor Bilder auf ungewöhnliche Merkmale, wie beispielsweise ein übermäßig perfektes Erscheinungsbild. Beide Detektoren vergeben eine Bewertung, welche darauf hinweist, ob das Bild KI-generiert sein könnte und werden regelmäßig Updates unterzogen. Sofern die Nutzer sich registriert haben können die Erkennungsfunktionen von Is It AI in eigene Anwendungen integriert werden. Diese erhalten nach Anmeldung Zugriff auf ein Token generierendes Dashboard.

AI or Not [12] erkennt KI-generierte Bilder, Texte, Musik, Deepfakes und Videos. Seine Erkennungsmodelle werden dabei regelmäßig aktualisiert. Auf seiner Webseite werben die Entwickler mit der vielseitigen Anwendbarkeit des Tools. Unter anderem soll es zum Aufdecken von gefälschten Propaganda-Inhalten, gefakten materiellen Schäden für Versicherungen sowie Deepfakes bei gefälschten Dokumenten mit personenbezogenen Daten wie Reisepässe oder Ähnliches genutzt werden können. Wie wahrheitsgemäß diese Aussagen ausfallen, können wir anhand unserer Evaluation nicht genau feststellen. Allerdings bestätigte sich die Aussage, dass AI or Not regelmäßig aktualisiert wird, da wir während unserer Evaluationsphase eine deutliche Veränderung im Design der Webseite feststellten.

Die Plattform **Hive Moderation** [13] gehört zur Firma The Hive und verlinkt teilweise auf dieselben Modelle: Die Modelle zur Inhaltsmoderation, die auch bei The Hive AI verlinkt werden, gehören zu Hive Moderation. The Hive hat verschiedene KI-Modelle, die sie zur

[10] https://illuminarty.ai/de/
[11] https://isitai.com/
[12] https://www.aiornot.com/
[13] https://hivemoderation.com/ai-generated-content-detection

Verfügung stellen. Dazu gehört auch die Kategorie Bilderkennung, welche nicht nur KI-generierte Inhalte herausfiltert, sondern in einer weiteren Analyse auch auf unangemessene Inhalte wie Nacktheit, Gewalt, Drogen und Hasssymbole hinweist. Letztere Funktion wurde jedoch nicht getestet.

4.3 Audio

Resemble.AI[14] ist ein Tool, welches KI-generierte Audios, insbesondere Deepfakes detektieren kann. Es richtet sich primär an Unternehmen und Institutionen, welche Audioinhalte auf ihre Authentizität hin prüfen möchten. Es existiert eine Demo-Version, die für alle kostenfrei zur Verfügung steht. In dieser Demo-Version muss eine E-Mail-Adresse angegeben werden, die einem drei freie Versuche ermöglicht. Dies kann man allerdings umgehen, indem man den Browserverlauf zurücksetzt. In dieser freien Version ist es möglich Dateien im .mp3- und .wav-Format mit einer Länge von bis zu einer Minute hochzuladen. Das Tool kann synthetische Medien innerhalb von 30 Millisekunden erkennen und ist somit für Echtzeitanwendungen geeignet.

Mithilfe des kostenlosen Webtools **ElevenLabs AI Speech Classifier** [15] , bei dem keine Anmeldung von Nöten ist, kann man prüfen, ob ein Audioclip mithilfe der von ElevenLabs entwickelten KI generiert wurde. Das Tool verwendet einen Klassifikator, der auf charakteristische Artefakte im synthetischen Audio trainiert ist. Dabei hat das Tool eine Präzision von >99% angegeben. Diese Präzision gilt allerdings nur für ein unverändertes, direkt heruntergeladenes ElevenLabs-Audio. Die Nutzung über den Browser ist intuitiv möglich. Laut eigener Angabe von ElevenLabs ist ihnen KI-Safety und Missbrauchsverhinderung sehr wichtig, weshalb dieser Detektor von ElevenLabs entwickelt worden ist, um Audios, die von Tools, welche von ElevenLabs entwickelt wurden, zu detektieren.

Die Plattform **Hive Moderation**[16] gehört zur Firma The Hive. The Hive hat verschiedene KI-Modelle, die sie zur Verfügung stellen. Hive Moderation bietet die Möglichkeit, die Ursprünge synthetischer Inhalte überprüfen zu können. Dabei werden Confidence Stores zurückgegeben, die angeben, wie wahrscheinlich es ist, dass der hochgeladene Inhalt KI-generiert ist. Folgende

[14] https://detect.resemble.ai/
[15] https://elevenlabs.io/de/ai-speech-classifier
[16] https://hivemoderation.com/ai-generated-content-detection

für das Medium “Audio” relevante Formate werden unterstützt: flac, mp3, mpeg, ogg, wav, x-wav, x-flac, x-m4a. Neben diesen Formaten kann Hive Moderation auch zur Erkennung von KI-generierten Bildern und Videos verwendet werden.

Deep Fake Total[17] ist ein von Fraunhofer AISEC entwickeltes Tool, um Deepfakes zu erkennen. Das Tool ist ein Open-Source Produkt, das allerdings keine industriellen Standards verwendet. Man kann entweder eigene Dateien hochladen und diese detektieren lassen oder aber URLs von YouTube bzw. Twitter/X einfügen. Das Tool analysiert diese Dateien auf mögliche KI-generierte Merkmale. Anschließend bekommt man eine Einordnung, ob die Audio-Datei KI-generiert ist oder nicht. Das Tool ist browserbasiert und kostenlos nutzbar. Man braucht keine Anmeldung durchzuführen. Deep Fake Total beinhaltet ebenfalls ein interaktives Online-Tool, mit dem man selbst ausprobieren kann, ob man KI-generierte Audio-Deepfakes erkennen kann.

4.4 Video

The Hive AI[18] ist eine Plattform, die neben Modellen zur Erkennung KI-generierter Inhalte ein breites Spektrum an KI-Modellen zur Verfügung stellt, z.B. Modelle zur Moderation von Inhalten, zur Erkennung von Personen und Logos oder zur Generierung von KI-Inhalten. Im Bereich des KI-Erkennungstools wird mit “human-level accuracy” und einem Set aus Trainingsdaten geworben, das aus Millionen von vielseitigen KI-generierten und menschlichen Bildern besteht, unter anderem aus Fotografien, digitaler und traditioneller Kunst und Memes. Bei der Analyse arbeitet das Tool mit einer Verlaufskurve, die u.a. zeigt, an welcher Stelle der eingegebene Inhalt mit welcher Wahrscheinlichkeit KI-generiert ist.

In kundengeführten Evaluationen soll das Tool besser abschneiden als andere Wettbewerber[19]. Das Tool ist zudem als Browsererweiterung nutzbar.

Die Plattform **Hive Moderation**[20] gehört ebenfalls zur Firma The Hive und verlinkt teilweise auf dieselben Modelle: Die Modelle zur Inhaltsmoderation, die auch bei The Hive AI verlinkt

[17] https://deepfake-total.com/
[18] https://thehive.ai/

[19] https://thehive.ai/apis/ai-generated-content-classification
[20] https://hivemoderation.com/

werden, gehören zu Hive Moderation. Bei der Erkennung KI-generierter Inhalte gibt es jeweils ein eigenes Modell. Das Modell hier analysiert die Inhalte getrennt nach Audio- und Videoinhalten und gibt jeweils eine Wahrscheinlichkeit an, wie sicher der Inhalt KI-generiert ist oder nicht. Die Plattform wirbt damit, dass sie mit ihren Produkten über 300 Plattformen unterstützen. Genannt werden z.B. Reddit und Bluesky.

Deepware.ai[21] ist eine Firma, die sich mit den Gefahren von Deepfakes auseinandersetzt und seit 2018 sowohl zur Erkennung als auch zur Generation von KI-Inhalten forscht. Zum Zeitpunkt der Testdurchläufe befindet sich das Tool zur Erkennung KI-generierter Inhalte noch in einer Betaversion. Neben einer Oberfläche, bei der sich entweder Videodateien hochladen oder Links zur Überprüfung auf KI einfügen lassen, gibt es weitere Ressourcen, die über die Entwicklung von Deepfakes aufklären. Das Ergebnis wird in Form eines Tachometers angezeigt und markiert den Inhalt als verdächtig, wenn ein Deepfake vermutet wird.

Cantilux[22] wirbt mit fortgeschrittener Erkennung KI-generierter Bilder und Videos, die zu detaillierten Ergebnissen führen soll. Es wird aufgeführt, dass die Erkennung zuverlässig ist, die Analyse schnell abläuft und hochgeladene Inhalte geschützt werden, indem sie nur lokal gespeichert werden. Nach Eingabe des Inhalts folgen eine Analyse von Mustern, Texturen und bestimmten Artefakten, die zu einer Gesamtwahrscheinlichkeit führen, ob KI vorliegt. Im Zeitraum der Testdurchläufe (Stand: Mai/Juni 2025) waren weder die Seite “About” noch die Seite “Contact” funktionsfähig, wodurch eine Kontaktaufnahme mit den Betreiber*innen der Website nicht möglich war. Insgesamt hätte der Detektor als Negativbeispiel dienen sollen, jedoch ist die Seite mittlerweile (Stand: Mitte August 2025) gar nicht mehr verfügbar oder auffindbar, weshalb der KI-Detektor nicht mehr genutzt werden kann.

Das **DeepFake-o-Meter**[23] ist ein Tool, welches Open-Source-Methoden anwendet, um KI-generierte Bilder, Videos und Audios zu erkennen. Für die Erkennung der Inhalte sind verschiedene Modelle zuständig, die jeweils auf verschiedene Arten von Deepfakes trainiert wurden. Vor dem Hochladen kann angegeben werden, ob man weiß, ob der Inhalt KI-generiert ist oder nicht und ob man einer Datenweitergabe an die Entwickler*innen zustimmt. Bei den

[21] https://scanner.deepware.ai/
[22] https://www.cantilux.com/
[23] https://zinc.cse.buffalo.edu/ubmdfl/deep-o-meter/landing_page

Erkennungsmethoden handelt es sich um Prototypen, die nach eigener Aussage noch zu fehlerhaften Ergebnissen führen können.

5. Kriterienkatalog

	Bewertung des Tools: **Name**	Trifft zu	Teils/Teils	Trifft gar nicht zu
	Das Tool...			
1	...ist so gestaltet, dass es intuitiv nutzbar ist			
2	...wirkt vertrauenswürdig, z.B. durch den Aufbau bzw. dessen Webseite und der Angabe einer Kontaktadresse			
3	...gibt Hintergrundinformationen zum Tool und Firma			
4	...schlüsselt auf, wie es bei der Auswertung eingegebener Daten vorgeht			
5	...behandelt persönliche, von Nutzenden eingegebenen Daten so, dass der Schutz personenbezogener Daten sichergestellt ist, z.B. kein (längerfristiges) Speichern eingegebener Daten			
6	...kommt zu einem richtigen Ergebnis			
7	...legt sinnvoll dar, wie das Ergebnis ermittelt wurde			
8	...verweist auf weitere Möglichkeiten und Angebote, mit denen Nutzende sich (selbstständig) informieren können, z.B. durch Tutorials			
9	...ist (auch ohne Anmeldung) kostenfrei nutzbar (ggf. mit Einschränkungen) oder hat eine Demo-Version, in der eine gewisse Anzahl an Freiversuchen möglich ist. Eine Bezahloption gibt es (mit Anmeldung), in der alle Funktionen enthalten wären.			
10	...macht Vorschläge, welche KI die vorliegenden Inhalte generiert haben könnte und gibt evtl. Wahrscheinlichkeiten an, wie sicher die Einschätzung ist			
11	...wird regelmäßig aktualisiert und erkennt neue Generationen von KI-Generatoren			
12	...schützt eingegebene Inhalte dem Urheberrecht entsprechend bzw. gibt Informationen zum Thema			
13	...reagiert und bearbeitet die Anfrage in einer angemessenen Zeit (<20 Sekunden)			
14	...ist barrierefrei zugänglich, sodass auch Menschen mit Behinderungen das Tool uneingeschränkt nutzen können (z. B. kompatibel mit Screenreadern, per Tastatur bedienbar, kontrastreiche Darstellung, einfache Sprache etc.)			

Dies ist ein Beispiel für einen Kriterienkatalog, mit dem KI-Detektoren-Tools evaluiert werden können. Je nach Tool und Medienart können Kriterien abgeändert bzw. hinzugefügt werden. Mit einer einfachen Checkliste wie dieser können Tools selbstständig überprüft werden.

Im Abbildungsverzeichnis sind leere Kriterienkataloge für alle Medientypen beigefügt. Diese können in bibliothekarischen Schulungsveranstaltungen für die Nutzer*innen bereitgestellt werden, um eine selbstständige Evaluation durchführen zu lassen.

6. Vorstellung idealtypisches Tool

Im folgenden Abschnitt wechselt der Fokus von den bisher vorhandenen Tools zu einem idealtypischen Tool, welches sowohl die bereits gesammelten Kriterien als auch medienspezifische "Bonus"-Funktionen erfüllen soll. Die Evaluation der gesichteten Tools hat gezeigt, dass bereits einige Tools existieren, die KI-generierte Inhalte erkennen (beziehungsweise dies behaupten), allerdings sind diese noch verbesserungsbedürftig. Aus

diesem Grund wird im Folgenden - basierend auf dem Kriterienkatalog - ein idealtypisches Tool beschrieben, dessen Funktionen sich als hilfreicher für die alltägliche und wissenschaftliche Nutzung erweisen würden.

Folgende Kriterien sollten unbedingt abgedeckt werden, sind jedoch in den bisher bestehenden Tools nicht zufriedenstellend durchgeführt worden:

1) **Zerlegung der Inhalte.** Das Tool kann selbstständig die Medientypen Text, Bild, Audio, Video zerlegen und als KI- bzw. menschengeneriert erkennen. Im Idealfall braucht man für unterschiedliche Medientypen keine vier verschiedenen Tools.

2) **Urheberrecht.** Die Website des Tools verfügt über detaillierte Informationen zum Thema Urheberrecht. Bisher wurde kein Tool gefunden, das darauf eingeht, ob und wie das Urheberrecht beim Scan der Inhalte sichergestellt wird. Nutzer*innen können auf diese Informationen zugreifen, ohne ein kostenpflichtiges Abonnement abschließen zu müssen.

3) **Barrierefreiheit.** Das Tool sollte für alle Personen trotz körperlicher oder geistiger Behinderung vollumfänglich nutzbar sein. Wichtige Punkte wären hier z.B. die Kompatibilität mit Screenreadern, der hohe Kontrast von Websites zur besseren Lesbarkeit, die Verfügbarkeit in leichter Sprache sowie allgemein die Möglichkeit, die Seite auf anderen Sprachen (als Englisch) zu nutzen.

4) **Kostenfreie Nutzung.** Das Tool ist dauerhaft kostenfrei nutzbar und hat keine limitierte Anzahl an Testversuchen. Dies ist vorteilhaft für Nutzergruppen wie z.B. Student*innen oder Schüler*innen. Zudem sollte jeder Mensch - unabhängig seiner finanziellen Situation - in der Lage sein, KI-generierte Inhalte detektieren zu können, da KI-Safety und Missbrauchsverhinderung in der heutigen Zeit immer wichtiger werden.

5) **Technische Abläufe.** Es muss transparent dargestellt werden, wie und mit welchen Daten die Tools trainiert werden und wie sie bei ihrer anschließenden Auswertung vorgehen. Das ist nötig, damit Nutzer*innen den Prozess besser nachvollziehen können, was einerseits Vertrauen in die Tools schafft und andererseits dazu beitragen könnte, dass sie solche Inhalte in Zukunft selbst besser erkennen können.

6) **Hinweis auf Fehlbarkeit von KI-Tools.** Das Tool sollte vor Nutzung einen Hinweis geben, dass der gescannte Inhalt auch intellektuell nochmal geprüft werden sollte. Ein KI-Tool ist nie unfehlbar.

7) **Hinweis auf kritischen Inhalt.** Das Tool sollte erkennen, ob der Inhalt, der hochgeladen wurde, aus historischer/gesellschaftlicher oder ähnlicher Sicht bedenklich ist und eine Warnung ausgeben. Das ist insofern relevant, als dass es bei der Internetsicherheit helfen würde.

8) **Mögliche Voreingenommenheit.** Manche Tools können KI-Inhalte nicht nur erkennen, sondern sie auch selbst erstellen. Diese Tools sollten transparent aufschlüsseln, ob dadurch bei der Erkennung der Inhalte oder bereits beim Training mögliche Interessenskonflikte bestehen, was die Zuverlässigkeit der KI-Erkennungstools angeht.

6.1 Text

Folgende Funktionen wären für ein textspezifisches Tool nützlich:

1) **Plagiatserkennung.** Das Tool sollte bei eingegebenen Texten sofort erkennen, falls es sich um ein Plagiat handelt und urheberrechtlich geschützte Werke eingescannt werden. Damit Nutzer*innen wissen, was genau plagiiert wird, sollen bibliographische Daten inklusive (Internet-)Quellen zum Ursprungswerk erscheinen. Außerdem wäre ein kurzer Infotext am Anhang des Scans ideal. Dort wird kurz erklärt, was genau ein Plagiat ist und wie man in so einem Fall vorgehen sollte.

2) **Hinweise zu Texten.** Auf der Website sollte erklärt werden, für welche Texte das Tool genutzt werden kann/soll. Es gibt einige Tools, die nicht auf den wissenschaftlichen Kontext ausgelegt sind und somit weniger gut Plagiate bzw. KI-generierte "wissenschaftliche" Inhalte erkennen. Mit dieser Art von Hinweis wüssten Nutzer*innen sofort, ob das Tool für sie in Frage kommt.

3) **Struktur von Text-Generatoren.** Bei einigen getesteten Tools wurde bei der Auswertung eine Satz- bzw. Phrasenanalyse aufgezeigt. Dort wurden vorliegende Wörter aufgelistet, welche die KI-Wahrscheinlichkeit erhöhen. Jedoch hat diese Auflistung in den meisten Fällen nicht viel Sinn gemacht, da teilweise Wörter wie "ist" als typisch KI markiert wurden. Der Gedanke hinter dieser Hilfestellung ist gut, allerdings muss diese Analyse optimiert werden. Interessant wäre zusätzlich eine Übersicht zu den bekanntesten KI-Generatoren und typische Strukturmerkmale, auf die man zugreifen kann, ohne selbst einen Text zu scannen. So können Nutzer*innen schnell einschätzen, wie aktuell das KI-Detektoren-Tool ist und ob verschiedene Generatoren erkannt werden.

6.2 Bild

Folgende Funktionen wären für ein bildspezifisches Tool nützlich:

1) **Hinweis auf Bilddaten.** Das Tool gibt an, welche Daten es zu dem hochgeladenen Bild ermitteln kann. Idealerweise kann es dazu auch Vergleiche mit Internetquellen extern ziehen.

2) **Bildanalyse.** Das Tool stellt dar, welche KI das Bild erstellt haben könnte. Idealerweise ist dazu keine kostenpflichtige Anmeldung nötig.

3) **Hinweis auf für eine KI typische Bildfehler.** Das Tool verweist auf irgendeine Weise (zum Beispiel durch eine farbige Markierung betroffener Stellen) in den hochgeladenen Medien auffällige Fehler, welche der KI beim Erstellen des Bildes unterlaufen sind. Anschließend erklärt es den Nutzern der Website, woran sie gefakte Inhalte erkennen können. Beispiel: "Dem Mann auf dem vorliegenden Foto fehlen 2 Finger. Derartige Fehler erscheinen bei KI-generierten Bildern häufig.", etc.

4) **Verweis auf Internetquellen.** Das Tool vermerkt - sofern vorhanden - tatsächliche Internetquellen, welche mit dem eingefügten Material übereinstimmen könnte, z.B.: Dieses Bild scheint mit folgenden Inhalten aus dem Internet übereinzustimmen: Nennen/ Aufzeigen der Quellen XYZ (funktionieren sollte das Ganze ähnlich wie eine Google-Rückwärtssuche). Die Ergebnisse sollen dabei helfen, die tatsächliche Inspiration für das gefakte Bild herauszufinden.

6.3 Audio

Für das Medium "Audio" wären folgende Funktionen nützlich:

1) **Inhaltserfassung.** Das Tool sollte sowohl erkennen, ob der Text, der in der Audio-Datei gesprochen oder gesungen wird, von einer KI generiert wurde, aber auch ob die Audio selbst, also die Stimme oder die Musik, von einer KI erstellt worden ist. Viele Tools können bisher "nur" die Audio selbst als KI-generiert erkennen, können aber den Inhalt (noch) nicht ausreichend erfassen.

2) **Transparenz.** Das Tool sollte darstellen können, was von wem erstellt worden ist. Hierzu wäre ein gemeinsames Tool, das alle Medien erkennen kann, von Vorteil, da hierbei sowohl der Text (in Audio-Form) detektiert werden könnte, als auch die Audio als solche. Diese Analyse sollte so dargestellt werden, dass man als Nutzer erkennt, woran das Tool erkennen konnte, dass die Datei von einer KI erstellt worden ist. Zum Beispiel könnte man aufzeigen lassen, dass an der Stelle x/Minute y die Stimme eine zu gleichmäßige Monotonie aufweist,

oder die Musik zu gleichförmig (zu wenige Höhen/Tiefen) besitzt, sodass man hier als Tool auch eine didaktische Arbeit leistet, nämlich der Frage, woran man KI-Audios erkennen kann.
3) **Medienspezifische Erkennung.** Das Tool sollte nicht nur Stimmen erfolgreich erkennen, sondern auch Musik und Soundeffekte sicher richtig detektieren. Im Moment scheint im Tool das Augenmerk auf Deepfake-Erkennung von Stimmen zu liegen, doch auch Musik und Lyrics als Gegenstand menschlicher Forschung sollten erkannt werden können.

6.4 Video

Folgende Funktionen wären für ein videospezifisches Tool nützlich:
1) **Detailliertere Analyse.** Das Tool sollte genau aufschlüsseln, an welchen Stellen im Video (möglicherweise) KI-generierte Inhalte vorliegen. Dafür wäre eine Verlaufskurve gut geeignet, die verschiedene Aspekte wie „KI-generiert", „nicht KI-generiert", „Deepfake" oder den Namen bestimmter Tools zur Erstellung KI-generierter Inhalte angibt. Zudem sollte angegeben werden, mit welcher Wahrscheinlichkeit der Inhalt wirklich KI-generiert ist bzw. wie sicher sich das Tool bei dem Ergebnis ist, um zu verhindern, dass Nutzende dem Tool vertrauen, ohne den Inhalt selbst eingeschätzt zu haben.
2) **Medienspezifische Erkennung.** Die Erkennung von KI-generierten Inhalten sollte nicht nur auf eine Art von KI-Videos begrenzt sein. Manche Tools wurden speziell für Deepfakes entwickelt und können diese mitunter gut erkennen, jedoch ist erstens nicht jedes gefälschte Video ein Deepfake, und zweitens ist die Einordnung, welcher Typ nun vorliegt, nicht ganz einfach. Die Nutzung „reiner" Deepfake-Tools birgt die Gefahr, dass Nutzende ein vermeintliches Deepfake in ein solches Tool hochladen, welches dann aber aufgrund fehlender Trainingsdaten nicht oder nicht richtig verarbeitet und erkannt wird.
3) **Separate Analyse von Video und Audio.** Viele KI-generierte Videos haben zusätzlich eine Audiospur. Bei der Analyse sollten die beiden Bereiche getrennt voneinander analysiert werden und abschließend erst zu einem Gesamtergebnis berechnet werden, da bei manchen KI-Videos zwar die Stimme oder die Musik im Hintergrund auch KI-generiert ist, es aber auch sein kann, dass bei einem echten Video eine KI-generierte Audiospur hinterlegt wurde. Auch möglich ist es z.B., dass die Stimme in einem Video echt ist und vertrauenswürdig erscheint, aber das Video nicht dazu passt, weil es mit KI erstellt wurde.

7. Chancen & Grenzen der Tool-Nutzung für Bibliotheken

Internetkompetenz/Digital Literacy.[24] Wer sich den potenziellen Gefahren der Künstlichen Intelligenz im Internet bewusst ist, dem bieten Erkennungstools eine gute Möglichkeit, sich sicherer im World Wide Web zu bewegen. Unter anderem unterstützen sie beispielsweise ihre Nutzer dabei, Falschnachrichten zu entlarven und sich gegen ihren Einfluss zu wappnen. Die Nutzer trainieren sich also eine gewisse Internetkompetenz an: Die Fähigkeit, erworbenes Wissen aus dem Internet nicht nur aufzunehmen, sondern auch damit umgehen zu können und es zu reflektieren.

Unterstützung der Inhaltsprüfung. Die Erkennungstools können ihren Nutzern potenziell dabei helfen, KI-generierte Inhalte selbst leichter zu erkennen. Je öfter sie sich mit den Tools auseinandersetzen, desto wahrscheinlicher ist es, dass sie mit der Zeit selbst Auffälligkeiten erkennen. Die intellektuelle Inhaltsprüfung und Erkennung von KI werden ebenfalls dadurch gestützt, dass die Tools noch nicht vollständig entwickelt sind und somit eine ausführliche Überprüfung der Scan-Ergebnisse sich oftmals noch als sinnvoll erweist.

Erweiterung von Schulungsangeboten. Gerade in der aktuellen Zeit zeigt die Künstliche Intelligenz im Internet immer mehr Präsenz. Gleichzeitig nehmen aber auch Überlegungen zu, wie mit diesem Fakt und der KI selbst fortan umgegangen werden sollte. Besonders der Fakt, dass die Entwicklung von intelligenten Computersystemen sich innerhalb weniger Jahre rasant und drastisch verbessert hat, regt zum Nachdenken an. Genau hier besteht für Bibliotheken die Möglichkeit, mit umfassenden Angeboten an das Thema anzuknüpfen. Verfügen die Mitarbeiter der Bibliotheken selbst über ein großes Wissen bezüglich Themen rund um die richtige KI-Nutzung und –Erkennung, so können sie dieses in bibliotheksnahen Angeboten (verschiedene Schulungen, Beratungsgespräche, etc.) an Nutzende weitergeben. Gerade solche, die wissenschaftliche Arbeiten schreiben, können darin womöglich einen zusätzlichen Anlass sehen, eine Bibliothek aufzusuchen und ihre Angebote zu nutzen.

Urheberrecht und Datenschutz. Urheberrechtsfragen stellen für Bibliotheken womöglich eine Grenze dar: Eine absolute Datensicherheit bei der Nutzung von KI-Erkennungstools ist nicht gegeben, da oft nicht explizit genannt wird, ob und inwiefern die Inhalte, die Nutzende in das

[24] https://wb-web.de/aktuelles/digital-literacy-versuch-einer-begriffsbestimmung.html

Tool eingeben, dem Urheberrecht entsprechend geschützt werden oder ob die Inhalte möglicherweise sogar als Trainingsdaten weitergenutzt werden.
Idealerweise werden Datenschutzrichtlinien (Privacy Policy) vor erstmaliger Nutzung auf derartige Informationen hin geprüft.
Es stellt sich dennoch die Frage, inwiefern die Tools in Schulungen integriert werden können, da häufig keine Aussage dazu getroffen wird, welche persönlichen Daten wie lange zu Verarbeitungs- und Analysezwecken gespeichert werden. Zudem stellen einige Tools nur einen begrenzten Umfang an Testversuchen zur Verfügung oder erfordern die Anmeldung mit einer Mailadresse.

Kosten und Anmeldung. Einige KI-Detektoren bieten Unternehmen und z.T. Privatpersonen die Möglichkeit an, eine kostenpflichtige Version mit umfangreicheren Funktionen zu lizenzieren. Einerseits ließ sich nicht herausfinden, ob Bibliotheken als Unternehmen gezählt werden, die eine solche Lizenz in Anspruch nehmen können, andererseits ist die Entwicklung der Detektoren noch nicht fortgeschritten und zuverlässig genug, als dass sich eine Lizenz lohnen würde. Da der Status der Bibliotheken als Unternehmen unklar ist, lassen sich auch keine Angaben machen, wie hoch die Kosten für eine Lizenz wären.
Falls Interesse an einer Lizenz besteht, sollen Bibliotheken prüfen, ob die Institution selbst (z.B. Universität) schon eine Lizenz hat und prüfen, ob eine bestimmte Anzahl an Lizenzierungen möglich ist.

Wandel der Tools im Laufe der Zeit. Derzeit entwickelt sich die generative KI sehr schnell. KI-Detektoren geben aber zum Teil nicht an, in welchen Abständen sie Aktualisierungen vornehmen, wodurch man nicht weiß, ob sie mit den Entwicklungen der Künstlichen Intelligenz mithalten können.

8. Fazit

Die Evaluation der vorliegenden KI-Detektoren-Tools hat gezeigt, dass der aktuelle Stand dieser verbesserungswürdig ist. Die Grundidee und Durchführung der Tools befindet sich noch in den Anfangsstadien, hat jedoch Entwicklungspotenzial. Die Nutzung der kostenpflichtigen Varianten der Tools lassen womöglich andere Schlussfolgerungen zu. Diese wurden allerdings

nicht getestet, da von der Sichtweise typischer Bibliothekszielgruppen (z.B. Student*innen, Schüler*innen) ausgegangen wurde, die eher kostenfreie Angebote der Bibliothek nutzen.

Das Ziel des Leitfadens war es, einen stichprobeartigen Überblick über die KI-Erkennungs-Tools zu geben und eine ungefähre Aussage zu deren aktuellen Stand zu treffen. Daher ist es wichtig zu betonen, dass die Testung der vorliegenden Tools nicht repräsentativ ist. Bestehende Tools werden kontinuierlich verbessert, was zur Folge hat, dass innerhalb kurzer Zeit unterschiedliche Schlussfolgerungen hinsichtlich der Funktionsfähigkeit und Zuverlässigkeit der Testergebnisse möglich sind. Diese Zuverlässigkeit ist derzeit jedoch ausbaufähig. Es besteht die Hoffnung, dass sich die vorhandenen Tools im Laufe der Zeit verbessern, um verschiedenen Nutzer*innengruppen als Hilfe zu dienen.

Aus diesem Grund ist neben der Benutzung von KI-Detektoren-Tools eine zusätzliche intellektuelle Prüfung ausschlaggebend. Da sich nicht nur die KI-Detektoren-Tools, sondern auch die generative KI weiterentwickelt, wird die intellektuelle Prüfung von KI-generierten Inhalten zunehmend schwieriger. Unabhängig von der Qualität der Tools sollten Nutzer*innen einen Mittelweg zwischen Toolnutzung und eigener Überprüfung finden. Nur so ist es möglich, fundierte Ergebnisse zu erzielen.

Basierend auf den Evaluationen unterscheidet sich vermutlich die Qualität der existierenden Tools. Eine Einbeziehung dieser Tools in bibliothekarischen Schulungen ist womöglich nur dann zielführend, wenn diese vorgestellt und Nutzer*innen nahegebracht werden, anstatt sie für wissenschaftliches Arbeiten zu empfehlen. Für diese Zwecke können Bibliotheken den beiliegenden Kriterienkatalog (vgl. Anhang) in ihren Veranstaltungen aushändigen, um Nutzer*innen die Chance zu geben, eigenständig Tools zu überprüfen. Da Informationskompetenz und in diesem Rahmen auf die Digital Literacy in Bibliotheken gefördert werden soll, könnte alternativ der Fokus von Schulungsveranstaltungen darauf ausgelegt werden, intellektuelle Prüfungsfähigkeiten zu stärken.

Unabhängig von der Nutzung von KI-Detektoren in Schulungen ist es für Bibliotheksmitarbeiter*innen relevant, sich mit diesem Thema auseinanderzusetzen, um Personen zu helfen, welche Unterstützung bei der Erkennung KI-generierter Inhalte benötigen. Um diesbezüglich eine gute erste Anlaufstelle für Nutzer*innen zu sein, müssen

Bibliotheksmitarbeiter*innen einen guten Kenntnisstand zu generativer KI in all ihren Formen haben, weshalb es wichtig ist, dass sie sich mit der Thematik beschäftigen.

9. Literaturverzeichnis

Kapitel 1:

Brockhaus Enzyklopädie Online, Generative KI (Informatik). https://brockhaus-1de-16i5jjdz50528.emedia1.bsb-muenchen.de/ecs/enzy/article/generative-ki-informatik (aufgerufen am 25.06.2025), NE GmbH Brockhaus

Brockhaus Enzyklopädie Online, Künstliche Intelligenz (Informatik). https://brockhaus-1de-16i5jjdz50528.emedia1.bsb-muenchen.de/ecs/enzy/article/kunstliche-intelligenz (aufgerufen am 25.06.2025), NE GmbH Brockhaus

Ladwig, P. (2024, Dezember 5). *Was ist KI und welche Formen von KI gibt es?* https://www.bpb.de/lernen/bewegtbild-und-politische-bildung/555997/was-ist-ki-und-welche-formen-von-ki-gibt-es/ (aufgerufen am 26.06.2025)

Sonar, T. (2023, August 29). *Was ist ein KI-Detektor? KI-Text erkennen und identifizieren*. https://www.typetone.ai/de/blog/ai-detectors-what-are-they-and-how-to-avoid-getting-detected

Sajid, H. (2024, November 20). *Understanding AI Detectors: How They Work and How to Outperform Them*. https://www.unite.ai/understanding-ai-detectors-how-they-work-and-how-to-outperform-them/ (aufgerufen am 27.06.2025)

Kapitel 3:

DW Innovation. (o. J.). *How to verify?* https://www.howtoverify.info/ (aufgerufen am 09.07.2025)

A-SIT Zentrum für sichere Informationstechnologie – Austria. (2025, März 29). *Täuschend echt: So erkennen Sie KI-Content*. https://www.onlinesicherheit.gv.at/Services/News/KI-Inhalte-erkennen.html

Gilbert, M. (2024, August 5). *#Faktenfuchs: So erkennen Sie KI-generierte Fakes*. https://www.br.de/nachrichten/netzwelt/ki-bilder-stimm-klone-ki-generierte-fakes-erkennen-faktenfuchs,UKCLt6o (aufgerufen am 09.07.2025)

Contentconsultants SEO Beratung. (o. J.). *KI-Texte erkennen: Typische Formulierungen und Muster*. https://www.contentconsultants.de/ki-texte-erkennen-warum-man-texte-besser-selbst-schreibt/ (aufgerufen am 26.06.2025)

Grammarly. (2025, April 7). *How Do AI Detectors Work? Key Methods, Accuracy, and Limitations*. https://www.grammarly.com/blog/ai/how-do-ai-detectors-work/#2 (aufgerufen am 26.06.2025)

Berliner Rundfunk. (o. J.). *KI-generierte Bilder erkennen: So entlarven Sie Deepfakes und Fälschungen*. https://www.berliner-rundfunk.de/so-entlarven-sie-deepfakes-und-faelschungen (aufgerufen am 27.06.2025)

MIT Media Lab. (o. J.). *Detect DeepFakes: How to counteract misinformation created by AI*. https://www.media.mit.edu/projects/detect-fakes/overview/ (aufgerufen am 27.06.2025)

Harding, X. (2024, Februar 26). *Wurde dieses Video mit KI generiert? So erkennen Sie es*. https://www.mozillafoundation.org/de/blog/sora-ai-video/ (aufgerufen am 09.07.2025)

Uhlenbrock, L. (2024). KI-Generierte Bilder, Texte und Videos erkennen. *merz - Zeitschrift für Medienpädagogik*, *68*(3). https://www.medienradar.de/hintergrundwissen/artikel/ki-generierte-bilder-texte-und-videos-erkennen (aufgerufen am 09.07.2025)

A-SIT Zentrum für sichere Informationstechnologie – Austria. (2023, Oktober 12). *Audio-Deepfakes und Voice-Cloning: So schützen Sie sich vor Betrug*. https://www.onlinesicherheit.gv.at/Services/News/Audio-Deepfake-Voice-Cloning.html (aufgerufen am 09.07.2025)

Sparrow, T. (2024, August 21). *Faktencheck: Wie erkenne ich Audio-Deepfakes?* https://www.dw.com/de/faktencheck-wie-erkenne-ich-audio-deepfakes/a-69980269 (aufgerufen am 10.07.2025)

ElevenLabs. (o. J.). https://elevenlabs.io/de/v3 (aufgerufen am 09.07.2025)

Kapitel 4:

https://app.gptzero.me/ (aufgerufen am 17.08.2025)

https://isgen.ai/de (aufgerufen am 17.08.2025)

https://www.zerogpt.com/ (aufgerufen am 17.08.2025)

https://smodin.io/de (aufgerufen am 17.08.2025)

https://quillbot.com/de/ (aufgerufen am 17.08.2025)

https://gptzero.notion.site/GPTZero-Release-Notes-Model-and-API 6f58686f6381498baef35212463b7da6 (aufgerufen am 26.06.2025)

https://gowinston.ai/de/ (aufgerufen am 09.08.2025)

https://illuminarty.ai/de/ (aufgerufen am 09.08.2025)

https://isitai.com/ (aufgerufen am 28.08.2025)

https://www.aiornot.com/ (aufgerufen am 09.08.2025)

https://hivemoderation.com/ai-generated-content-detection(aufgerufen am 09.08.2025)

https://detect.resemble.ai/ (aufgerufen am 31.07.2025)

https://elevenlabs.io/de/ai-speech-classifier (aufgerufen am 31.07.2025)

https://hivemoderation.com/ai-generated-content-detection (aufgerufen am 31.07.2025)

https://deepfake-total.com/ (aufgerufen am 31.07.2025)

https://thehive.ai/ (aufgerufen am 09.07.2025)

https://thehive.ai/apis/ai-generated-content-classification (aufgerufen am 09.07.2025)

https://hivemoderation.com/ (aufgerufen am 12.07.2025)

https://scanner.deepware.ai/ (aufgerufen am 12.07.2025)

https://www.cantilux.com/ (aufgerufen am 12.07.2025)

https://zinc.cse.buffalo.edu/ubmdfl/deep-o-meter/landing_page (aufgerufen am 12.07.2025)

Kapitel 7:

Kilian, L. (2019, Oktober 10). *„Digital Literacy“—Versuch einer Begriffsbestimmung*. https://wb-web.de/aktuelles/digital-literacy-versuch-einer-begriffsbestimmung.html (aufgerufen am 30.06.2025)

10. Abbildungsverzeichnis

	Bewertung des Tools (Text):	Trifft zu	Teils/ Teils	Trifft gar nicht zu	Bemerkungen
	Das Tool...				
1	...ist so gestaltet, dass es intuitiv nutzbar ist				
2	...wirkt vertrauenswürdig, z.B. durch den Aufbau bzw. dessen Webseite und der Angabe einer Kontaktadresse				
3	...schlüsselt auf, wie es bei der Auswertung eingegebener Daten vorgeht				
4	...behandelt persönliche, von Nutzenden eingegebenen Daten so, dass der Schutz personenbezogener Daten sichergestellt ist, z.B. kein (längerfristiges) Speichern eingegebener Daten				
5	...kommt zu einem richtigen Ergebnis				
6	...legt sinnvoll dar, wie das Ergebnis ermittelt wurde				
7	...verweist auf weitere Möglichkeiten und Angebote, mit denen Nutzende sich (selbstständig) informieren können, z.B. durch Tutorials				
8	...ist (auch ohne Anmeldung) kostenfrei nutzbar (ggf. mit Einschränkungen) oder hat eine Demo-Version, in der eine gewisse Anzahl an Freiversuchen möglich ist. Eine Bezahloption gibt es (mit Anmeldung), in der alle Funktionen enthalten wären.				
9	...macht Vorschläge, welche KI die vorliegenden Inhalte generiert haben könnte und gibt evtl. Wahrscheinlichkeiten an, wie sicher die Einschätzung ist				
10	...wird regelmäßig aktualisiert und erkennt neue Generationen von KI-Generatoren				
11	...schützt eingegebene Inhalte dem Urheberrecht entsprechend bzw. gibt Informationen zum Thema				
12	...reagiert und bearbeitet die Anfrage in einer angemessenen Zeit (<20 Sekunden)				
13	...ist barrierefrei zugänglich, sodass auch Menschen mit Behinderungen das Tool uneingeschränkt nutzen können (z. B. kompatibel mit Screenreadern, per Tastatur bedienbar, kontrastreiche Darstellung, einfache Sprache etc.)				

	Bewertung des Tools (Bild):	Trifft zu	Teils/ Teils	Trifft gar nicht zu	Bemerkungen
	Das Tool...				
1	...ist so gestaltet, dass es intuitiv nutzbar ist.				
2	...wirkt vertrauenswürdig, z.B. durch den Aufbau bzw. dessen Webseite und der Angabe einer Kontaktadresse.				
3	...schlüsselt auf, wie es bei der Auswertung eingegebener Daten vorgeht.				
4	...behandelt persönliche, von Nutzenden eingegebenen Daten so, dass der Schutz personenbezogener Daten sichergestellt ist, z.B. kein (längerfristiges) Speichern eingegebener Daten.				
5	...kommt zu einem richtigen Ergebnis.				
6	...legt sinnvoll dar, wie das Ergebnis ermittelt wurde.				
7	...verweist auf weitere Möglichkeiten und Angebote, mit denen Nutzende sich (selbstständig) informieren können, z.B. durch Tutorials.				
8	...erkennt vielfältige Medien als KI- bzw. nicht KI-generiert.				
9	...ist (auch ohne Anmeldung) kostenfrei nutzbar (ggf. mit Einschränkungen) oder hat eine Demo-Version, in der eine gewisse Anzahl an Freiversuchen möglich ist. Eine Bezahloption gibt es (mit Anmeldung), in der alle Funktionen enthalten wären.				
10	...macht Vorschläge, welche KI die vorliegenden Inhalte generiert haben könnte und gibt evtl. Wahrscheinlichkeiten an, wie sicher die Einschätzung ist.				
11	...wird regelmäßig aktualisiert und erkennt neue Generationen von KI-Generatoren.				
12	...schützt eingegebene Inhalte dem Urheberrecht entsprechend bzw. gibt Informationen zum Thema				
13	...reagiert und bearbeitet die Anfrage in einer angemessenen Zeit (<20 Sekunden)				
14	...ist barrierefrei zugänglich, sodass auch Menschen mit Behinderungen das Tool uneingeschränkt nutzen können (z. B. kompatibel mit Screenreadern, per Tastatur/Sprache bedienbar, kontrastreiche Darstellung, einfache Sprache etc.).				

Abbildung 3: Kriterienkatalog für den Medientyp Audio

	Bewertung des Tools (Audio):	Trifft zu	Teils/ Teils	Trifft gar nicht zu	Bemerkungen
	Das Tool...				
1	... ist so gestaltet, dass es intuitiv benutzbar ist				
2	...wirkt vertrauenswürdig, z.B. durch den Aufbau bzw. dessen Webseite und der Angabe einer Kontaktadresse.				
3	...schlüsselt auf, wie es bei der Auswertung eingegebener Daten vorgeht.				
4	...behandelt persönliche, von Nutzenden eingegebene Daten so, dass der Schutz personenbezogener Daten sichergestellt ist, z.B. kein (längerfristiges) Speichern eingegebener Daten.				
5	...kommt zu einem richtigen Ergebnis.				
6	... legt sinnvoll dar, wie das Ergebnis ermittelt wurde.				
7	...verweist auf weitere Möglichkeiten und Angebote, mit denen Nutzende sich (selbstständig) informieren können, z.B. durch Tutorials oder einen Chatbot				
8	...erkennt KI-generierte Musik genauso gut, wie verschiedene Stimmen mit verschiedenen Sprachen und Stimmlagen				
8,1	bei Musik-Audio: ...erkennt einzelne Genres, erkennt Instrumentalmusik gleich zuverlässig wie Gesang				
8,2	bei Stimmen: ...erkennt verschiedene Sprachen, Stimmlagen, etc.				
9	...ist (auch ohne Anmeldung) kostenfrei nutzbar (ggf. mit Einschränkungen) oder hat eine Demo-Version, in der eine gewisse Anzahl an Freiversuchen möglich ist. Eine Bezahloption gibt es (mit Anmeldung), in der alle Funktionen enthalten wären.				
10	...macht Vorschläge, welche KI die vorliegenden Inhalte generiert haben könnte und gibt evtl. Wahrscheinlichkeiten an, wie sicher die Einschätzung ist.				
11	... wird regelmäßig aktualisiert und erkennt neue Generationen von KI-Generatoren.				
12	...schützt eingegebene Inhalte dem Urheberrecht entsprechend bzw. gibt Informationen zum Thema				
13	...reagiert und bearbeitet die Anfrage in einer angemessenen Zeit. (<1 Minute)				
14	...ist barrierefrei zugänglich, sodass auch Menschen mit Behinderungen das Tool uneingeschränkt nutzen können (z. B. kompatibel mit Screenreadern, per Tastatur/Sprache bedienbar, kontrastreiche Darstellung, einfache Sprache etc.).				

Abbildung 4: Kriterienkatalog für den Medientyp Video:

	Bewertung des Tools (Video):	Trifft zu	Teils/ Teils	Trifft gar nicht zu	Bemerkungen
	Das Tool...				
1	.ist so gestaltet, dass es intuitiv nutzbar ist.				
2	...wirkt vertrauenswürdig, z.B. durch den Aufbau bzw. dessen Webseite und der Angabe einer Kontaktadresse.				
3	..schlüsselt auf, wie es bei der Auswertung eingegebener Daten vorgeht.				
4	personenbezogener Daten sichergestellt ist, z.B. kein (längerfristiges) Speichern				
5	...kommt zu einem richtigen Ergebnis.				
6	... legt sinnvoll dar, wie das Ergebnis ermittelt wurde.				
7	...verweist auf weitere Möglichkeiten und Angebote, mit denen Nutzende sich (selbstständig) informieren können, z.B. durch Tutorials				
8	...erkennt vielfältige Videos als KI bzw. nicht-KI (verschiedene Sprachen/Stimmlagen, menschliche Anpassungen, Deepfakes, verschiedene Stile: realistisch, animiert...)				
9	...ist (auch ohne Anmeldung) kostenfrei nutzbar (ggf. mit Einschränkungen) oder hat eine Demo-Version, in der eine gewisse Anzahl an Freiversuchen möglich ist. Eine Bezahloption gibt es (mit Anmeldung), in der alle Funktionen enthalten wären				
10	...macht Vorschläge, welche KI die vorliegenden Inhalte generiert haben könnte und gibt evtl. Wahrscheinlichkeiten an, wie sicher die Einschätzung ist.				
11	... wird regelmäßig aktualisiert und erkennt neue Generationen von KI-Generatoren.				
12	...schützt eingegebene Inhalte dem Urheberrecht entsprechend bzw. gibt Informationen zum Thema				
13	...reagiert und bearbeitet die Anfrage in einer angemessenen Zeit. (<1 Minute)				
14	uneingeschränkt nutzen können (z. B. kompatibel mit Screenreadern, per Tastatur bedienbar, kontrastreiche Darstellung, einfache Sprache etc.).				