Technische Hochschule Ostwestfalen-Lippe

Fachbereich Medienproduktion

Bachelorarbeit

KI-generierte Stimmen in unterschiedlichen Nutzungskontexten auf YouTube:

Eine explorative, vergleichende Analyse von YouTube-Kommentaren

von

Tobias Cornelsen

gemäß Bachelorprüfungsordnung für den Studiengang Medienproduktion (BPO MP 2018) in der Fassung der Bekanntmachung vom 26. Mai 2020

1. Prüfer: Prof. Dipl.-Des. Paul Steinmann
2. Prüfer: Prof. Dr. phil. Tobias Schmohl

Die Bachelorarbeit umfasst 59 Seiten.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Abkürzungsverzeichnis

AI – Authentizität und Identität (Hauptkategorie)

API – Application Programming Interface

CSV – Comma-Separated Values (Dateityp)

GV – Glaubwürdigkeit und Vertrauen (Hauptkategorie)

KI – Künstliche Intelligenz

MOS – Mean Opinion Score

NK – Natürlichkeit und Klang (Hauptkategorie)

NR – Nicht relevant (Restkategorie)

RK – Reaktion auf KI-Einsatz (Hauptkategorie)

TSV – Tab-Separated Values (Dateityp)

TTS – Text-to-Speech

VTuber – Virtual YouTuber

1 Einleitung

1.1 Problemfeld und Relevanz

Mit den rasanten Entwicklungen der letzten Jahre im Bereich künstlicher Intelligenz haben auch synthetische Stimmen einen Qualitätssprung gemacht (Orynbay et al., 2024). KI-basierte neuronale Text-to-Speech-Systeme erreichen erstmals hinreichende Menschenähnlichkeit (Nussbaum et al., 2025), das zeigt auch eine Untersuchung von Bruder et al. (2025, S. 5), in der KI-Stimmen nur in etwa 55 % der Fälle korrekt als synthetisch und nicht menschlich identifiziert wurden. Durch Plattformen wie ElevenLabs ist die Nutzung KI-generierter Stimmen inzwischen auch für Medienproduzierende leicht zugänglich (ElevenLabs, 2024). Dadurch eröffnen sich neue Möglichkeiten für Gestaltung und Produktion, aber auch wirtschaftliche Überlegungen im Bereich audiovisueller Medien. Die Wahl einer KI-Stimme ist eine wichtige gestalterische Entscheidung. Nach Chion (1999) beeinflusst die Positionierung im audiovisuellen Raum die Wahrnehmung. Dazu belegen Cambre & Kulkarni (2019), dass Stimmen unabhängig vom Inhalt soziale Signale übertragen und folglich ein wichtiger Faktor in der Gestaltung sind. Ebenso wichtig für die Wahrnehmung ist der Verwendungskontext, gerade für die Glaubwürdigkeit (Maltezou-Papastylianou et al., 2025b; Schreibelmayr & Mara, 2022).

Für Medienproduzierende ist YouTube als größte Videoplattform mit über 2,5 Milliarden monatlichen Nutzern (We Are Social et al., 2025) ein zentraler Ort für Veröffentlichung, Verbreitung und Rezeption audiovisueller Inhalte. Auf YouTube setzen bereits viele Kanäle KI in der Videoproduktion (Metricool, 2025). Eine Umfrage von WeCreate (2025) zeigte jedoch auch, dass die Akzeptanz von KI-Nutzung bei über zwei Dritteln der Menschen zwischen 16 und 39 Jahren in Deutschland selbst bei gutem Inhalt gering ausfällt. Kommentare zeigen die Wahrnehmung von Kommentierenden in einem realen Umfeld und sind eine valide Basis für die Rezeptionsforschung und Erfassung der öffentlichen Aushandlung (Thelwall, 2018).

1.2 Forschung

Die Forschung zeigt, dass Natürlichkeit als stärkster Prädiktor der Gesamtbewertung synthetischer Stimmen wirkt (Hinterleitner, 2017, S. 49). Auch die Kontextabhängigkeit der Akzeptanz (Schreibelmayr & Mara, 2022, S. 9) und der Einfluss von Wissen über die KI-Herkunft (Fan & Liu, 2025, S. 2060) konnten bereits belegt werden. Dies legt nahe, dass die Funktion der Stimme einen Einfluss auf die Wahrnehmung hat.

Alle genannten Studien arbeiten unter Laborbedingungen und bisher noch nicht im Plattformkontext von YouTube. Cooper et al. (2024, S. 3) beschreiben, dass Ergebnisse bestehender Forschung aufgrund dieser Variable nicht verglichen werden können. Diese Arbeit analysiert erstmals explorativ die funktionsspezifische Wahrnehmung synthetischer Stimmen in

YouTube-Kommentaren. Bisherige Forschung dieser Art fand ausschließlich unter Laborbedingungen statt und spiegelt demnach nicht die Reaktionen in einem Kontext wider, der direkt für die Medienproduktion relevant ist. Diese Arbeit schließt diese Lücke nicht, kann aber zeigen, welche Kategorien und Muster plattformspezifische Forschung braucht, um die Wahrnehmung von Reaktionen auf KI-Stimmen auf YouTube zu verstehen.

Ziel der Arbeit ist es, erste Hypothesen darüber aufzustellen, wie Medienproduzierende KI-Stimmen auf YouTube nutzen können und was dabei zu beachten ist. Sie soll zur ersten Orientierung zeigen, welche Faktoren bei KI-Stimmen-Nutzung auf YouTube relevant sein könnten. Die Arbeit kann keine allgemeingültigen Aussagen und Kausalaussagen leisten und hat keine Repräsentativitätsansprüche.

1.3 Forschungsfragen

1.3.1 Vorannahme

Die Untersuchung geht mit der Vorannahme ins Feld, dass die Wahrnehmung und Bewertung synthetischer Stimmen funktionsabhängig variieren. Diese Annahme wird aufgrund der explorativen Form der Arbeit nicht als zu prüfende Hypothese verstanden, sondern als Ausgangspunkt für die explorative Analyse.

1.3.2 Hauptfrage

Wie werden KI-generierte Stimmen in YouTube-Kommentaren in Abhängigkeit von ihrer Funktion im Video wahrgenommen und bewertet?

1.3.3 Unterfragen

Welche Formen von Authentizität und fehlender Authentizität werden Kommentaren zufolge den Stimmen zugeschrieben?

Welche Bewertungsdimensionen sind funktionsspezifisch und welche treten funktionsübergreifend auf?

Welche produktionsbezogenen Konsequenzen lassen sich aus den funktionsspezifischen Mustern ableiten?

1.4 Untersuchungszugriff & Aufbau

Es wird mit einem explorativen Design nach Stebbins (2001) gearbeitet, um Hypothesen zu generieren. Als Daten dienen Kommentare, die durch eine qualitative Inhaltsanalyse nach Kuckartz (2016) ausgewertet werden. Aus der Forschung werden Hauptkategorien gebildet und aus dem Material heraus Unterkategorien. Kommentare werden als Daten genutzt, da sie als einzige Datenform Plattformkontext und reale Rezeptionssituation abbilden.

Im Anschluss an die Einleitung wird der Forschungsstand zu synthetischen Stimmen, Authentizität auf YouTube und Kommentaren als Datenbasis erarbeitet. Daran schließt die detaillierte Beschreibung der Methodik mit Fallauswahl, Datenerhebung und Kodierlogik an. Als Nächstes werden die Ergebnisse festgehalten und beschrieben, um sie im Anschluss in der Diskussion zu interpretieren und einen Ausblick zu geben.

2 Forschungsstand und analytischer Rahmen

2.1 Synthetische Stimmen in medialer Produktion

Der Ausdruck „KI-generierte Stimmen“ bezeichnet in dieser Arbeit Stimmen, die durch synthetische Spracherzeugung auf Basis neuronaler Netzwerke erzeugt wurden. Im Folgenden werden diese vereinfachend als „synthetische Stimmen“ bezeichnet. Ning et al. (2019, S. 1) bezeichnen Sprachsynthese beziehungsweise Text-to-Speech (TTS) als Umwandlung von Texteingabe in Sprachausgabe. In der Vergangenheit dominierten Verfahren wie die Verkettung von Wellenformen (Concatenative Speech Synthesis) und parametrische Sprachsynthese, bei der stimmliche Eigenschaften mathematisch modelliert werden und aus diesen Parametern ein Sprachsignal berechnet wird (Ning et al., 2019, S. 3). Die in dieser Arbeit betrachteten synthetischen Stimmen basieren hingegen auf neuronalen Netzwerken, da diese wesentliche Vorteile, darunter bessere Natürlichkeit und weniger manuelle Vorarbeit, bieten (Orynbay et al., 2024, S. 11; Xie et al., 2025, S. 21,23). Da im Zentrum dieser Arbeit die Rezeption synthetischer Stimmen in verschiedenen Funktionen und die Bedeutung für die Produktion dieser Inhalte steht, sind vor allem Verfahren relevant, welche eine hinreichende perzeptuelle Menschenähnlichkeit erreichen. Nach Nussbaum et al. (2025, S. 472–473) stellt die Wahrnehmung von Menschenähnlichkeit einen zentralen Aspekt der Stimm-Natürlichkeit dar. Da insbesondere neuere, auf neuronalen Netzwerken basierende Text-to-Speech-Verfahren in der Lage sind, diese Menschenähnlichkeit überzeugend zu erzielen (Orynbay et al., 2024, S. 11; Xie et al., 2025, S. 11), fokussiert sich die vorliegende Analyse auf diese Technologie.

Für die Untersuchung sind zwei Gruppen innerhalb der neuronalen Text-to-Speech-Verfahren relevant. Die erste Gruppe sind generische Text-to-Speech-Stimmen. Hierbei handelt es sich um synthetische Stimmen, die als anonyme Erzähler fungieren und keine reale Identität beanspruchen. Bei der Rezeption dieser Stimmen liegt der Fokus des Publikums primär auf der Wahrnehmungsdimension der Natürlichkeit (Naturalness) (Nussbaum et al., 2025). Die zweite Gruppe umfasst das Stimmklonen (Voice Cloning). Hierbei werden gezielt Identität, Stimmfarbe und Betonung einer bestimmten, oft real existierenden Person repliziert (Alali & Theodorakopoulos, 2025; Azzuni & Saddik, 2025). Hierbei verschiebt sich die Wahrnehmungsdimension von der Natürlichkeit hin zur Authentizität. In den Kommentaren solcher Videos äußert sich die Verunsicherung darüber, auditiven Beweisen nicht mehr trauen zu können und eine kritische Hinterfragung der Intentionen der Creator (Çalli & Alma Çalli, 2025). Diese Form ist auch deshalb relevant, weil geklonte Stimmen bekannter Personen bei Rezipienten und Rezipientinnen parasoziale Beziehungen stärken können, aber auch Abwehrreaktionen erzeugen, wenn sie als Täuschung wahrgenommen werden (Kang et al., 2025). Darüber hinaus muss bei der Analyse der Kommentare berücksichtigt werden, dass moderne synthetische Stimmen häufig eine gezielte Steuerung von Emotionen und Sprechstilen erlauben (Xie et al., 2025).

Ob eine Stimme vom Publikum als passend oder deplatziert empfunden wird, hängt maßgeblich davon ab, ob die künstliche Expressivität als angemessen bewertet oder den Eindruck der „Künstlichkeit“ noch verstärkt (Çalli & Alma Çalli, 2025; Kang et al., 2025; Schreibelmayr & Mara, 2022). Um zu verstehen, wie diese Stimmen wahrgenommen werden, werden im Folgenden die relevanten Bewertungsdimensionen aus der Forschung erarbeitet.

2.2 Wahrnehmung synthetischer Stimmen

2.2.1 Klassische Qualitätsdimension

Historisch ist der Mean Opinion Score (MOS) das Standardmaß der Bewertung synthetischer Stimmen gewesen, jedoch ist dieser Wert bei verschiedenen Studien nicht vergleichbar und die Bewertung nach Natürlichkeit subjektiv, da kein universelles Verständnis davon existiert (Cooper et al., 2024). Deshalb braucht es zunehmend kontextspezifische Evaluationsansätze, die Nuancen wie Prosodie und emotionalen Ausdruck besser abbilden.

Hinterleitner (2017, S. 66) leitete aus verschiedenen Hörstudien ab, dass es fünf universelle Qualitätsdimensionen zur Bewertung synthetischer Stimmen gibt: Natürlichkeit der Stimme, prosodische Qualität, Flüssigkeit und Verständlichkeit, Störungsfreiheit und Gelassenheit. Dabei erwies sich „Natürlichkeit“ als stärkster Prädiktor für die Gesamtqualitätsbewertung. Anders als der MOS wird hier erstmals die technische Verständlichkeit und wahrgenommene Qualität getrennt betrachtet und nicht in einer Zahl zusammengefasst (Hinterleitner, 2017).

2.2.2 Erweiterte Dimensionen und soziale Wirkung

Kühne et al. (2020, S. 4) erweiterten in ihrer Studie die fünf Dimensionen nach Hinterleitner. Darin nutzten sie Verständlichkeit (intelligibility), Prosodie (prosody), Glaubwürdigkeit (trustworthiness), Selbstsicherheit (confidence), Begeisterung (enthusiasm), Angenehmheit (pleasantness), Menschlichkeit (human-likeness), Sympathie (likability) und Natürlichkeit (naturalness). Anhand dieser wurden die Stimmen aus drei Kategorien, synthetisch, humanoid und menschlich, bewertet. Parallel wurden die Sprecher nach den Dimensionen Anziehungskraft (appeal), Glaubwürdigkeit (credibility), Menschlichkeit (human-likeness) und Unheimlichkeit (eeriness) bewertet. Für die Stimmen waren die Angenehmheit und Glaubwürdigkeit signifikante Prädiktoren für deren wahrgenommene Sympathie (Kühne et al., 2020, S. 4, 7). Schreibelmayr & Mara (2022) ergänzen die Analyse um die Kontextdimension. Demnach variiert die Akzeptanz synthetischer Stimmen je nach Verwendungskontext. Für soziale Anwendungskontexte wurde insgesamt eine niedrigere Akzeptanz gegenüber synthetischen Stimmen festgestellt und eine signifikant höhere gegenüber der menschenähnlichsten Stimme (Schreibelmayr & Mara, 2022, S. 9). Ein weiter differenziertes Dimensionsmodell für die vergleichende Bewertung von KI- und menschlichen Stimmen bieten Chen et al. (2025, S. 12, 15)

in ihrer Studie. Da es sich hierbei um einen Preprint handelt, muss es jedoch mit Vorsicht betrachtet werden. In der Studie wurden menschliche Stimmen in den zwei Kerndimensionen „Soziale Anziehungskraft“ (Social Appeal) und „Stimmlicher Ausdruck“ (Vocal Expressiveness) konsistent besser bewertet. Diese Dimensionen sind auch in der vorliegenden Arbeit wichtig, um zu verstehen, warum Zuschauer synthetische Stimmen negativ bewerten. Bei objektiv ausdrucksstarker Prosodie wird bei Hörenden keine äquivalente soziale Wirkung erzielt, wenn sie wissen, dass es sich um eine KI-Stimme handelt (Chen et al., 2025, S. 11). Trotz stetigem Fortschritt im Bereich der neuronalen Stimmsynthese (Orynbay et al., 2024) können synthetische Stimmen oft noch nicht, wie menschliche Stimmen, durch Prosodie linguistische und paralinguistische Informationen wie emotionale Zustände des Sprechers völlig natürlich übertragen. Spontan produzierte Sprache klingt messbar anders als gespielte und insbesondere synthetischen Stimmen mangelt es in diesem Kontext noch an Ausdrucksauthentizität (Larrouy-Maestri et al., 2025, S. 29–30).

Nussbaum et al. (2025, S. 476) stellen eine Kontextabhängigkeit der wahrgenommenen Natürlichkeit fest, so kann dieselbe Stimme in einem Kontext natürlicher wirken als in einem anderen. Natürlichkeit korreliert außerdem stark mit Vertrauenswürdigkeit, Angenehmheit und Glaubwürdigkeit und zeigt, dass die Dimensionen gegenseitigen Einfluss haben (Kühne et al., 2020).

Stern et al. (2006) stellten einen Zusammenhang zwischen den Präferenzen für ein bestimmtes Sprachsynthesesystem und dessen Verständlichkeit fest. Dadurch kann darauf geschlossen werden, dass die Dimension der Verständlichkeit auch mit den Hörpräferenzen und nicht ausschließlich objektiv gemessen werden kann. Auch beschreibt die Studie, dass Verständlichkeit die Basisvoraussetzung zur Beurteilung ist (Stern et al., 2006). Taake (2009) stellte fest, dass bei ähnlichen Verständlichkeitswerten synthetische Sprache gegenüber natürlicher Sprache einen höheren kognitiven Verarbeitungsaufwand benötigt. Dieser Faktor wird bei simpler Einteilung nach Verständlichkeit vernachlässigt. Da es sich bei der Quelle um eine Dissertation handelt, werden die Befunde mit gebotener Vorsicht behandelt. Dennoch ist diese Art der Betrachtung wichtig, um in der vorliegenden Arbeit nachzuvollziehen, wie sich Bewertungen in Kommentaren bilden.

Auch bei der Glaubwürdigkeit synthetischer Stimmen gibt es Hinweise, dass diese kontextabhängig ist (Maltezou-Papastylianou et al., 2025b, S. 14). Während Zusammenfassungen früherer Studien teilweise nahelegen, dass der Einsatz von KI bei objektiven, neutralen Sachinformationen nicht zwingend zu einem Glaubwürdigkeitsverlust führt (Gong, 2023, S. 2), zeigen die eigenen experimentellen Untersuchungen von Gong (2023, S. 6, 11), dass Zuhörende KI-generierten Stimmen in Nachrichtenformaten selbst bei neutralen Inhalten insgesamt mit größerer Skepsis bezüglich ihrer Glaubwürdigkeit begegnen als menschlichen Stimmen. Diese Kontextabhängigkeit lässt sich in der vorliegenden Arbeit auch auf die

Funktionskategorien übertragen. Das wird durch Voorveld et al. (2025, S. 2925) bestärkt, die betonen, dass nicht nur der reine Klang einer synthetischen Stimme, sondern maßgeblich auch die Erwartungshaltung der Nutzenden die Reaktionen auf den Sprachagenten formt. Zudem gilt die wahrgenommene Menschenähnlichkeit laut ihnen als zentraler Einflussfaktor auf das Vertrauen und die Glaubwürdigkeit (Voorveld et al., 2025, S. 2917). In einer Studie von Maltezou-Papastylianou et al. (2025b, S. 9) wurden synthetische Stimmen bei neutralem Tonfall als vertrauenswürdiger als menschliche Stimmen eingestuft. Wenn menschliche Sprecher jedoch bewusst versuchten vertrauenswürdig zu klingen, übertrafen sie die synthetischen Stimmen. Soziale Urteile über synthetische Stimmen werden aufgrund derselben Grundstrukturen wie bei menschlichen Stimmen gemacht. Valenz ist eng mit der Vertrauenswürdigkeit und Dominanz mit Kompetenz verbunden (Shiramizu et al., 2022, S. 3).

Freundlichkeit (Pleasantness) und Vertrauenswürdigkeit (Trustworthiness) sind nach Kühne et al. (2020, S. 7, 11) die stärksten akustischen Prädiktoren für Sympathie (Likability) und nicht Natürlichkeit allein. Auch hier zeigt sich wieder, wie die verschiedenen Dimensionen miteinander zusammenhängen. Menschliche Stimmen werden als sympathischer und ausdruckstärker eingestuft (Kühne et al., 2020, S. 10).

2.2.3 Uncanny Valley bei synthetischen Stimmen

Das Uncanny Valley beschreibt den Effekt, dass künstliche, fast menschlich wirkende Figuren oft als besonders unheimlich wahrgenommen werden, gerade im Vergleich zu weniger menschlich wirkenden Figuren (Mori et al., 2012). Diesen Effekt vermuteten viele auch bei künstlichen (synthetischen) Stimmen. Entgegen der Erwartungen haben verschiedene Studien bisher jedoch kein allgemeines Uncanny Valley bei synthetischen Stimmen feststellen können (Baird et al., 2018; Kühne et al., 2020; Roesler et al., 2021; Romportl, 2014; Schreibelmayr & Mara, 2022). Synthetische Stimmen werden umso besser bewertet, je menschenähnlicher sie klingen. Im Kontext der Dimensionen werden sie als weniger unheimlich und sympathischer bewertet (Kühne et al., 2020, S. 11).

2.2.4 Zuschreibung synthetischer Stimmen

Als Basis, um zu verstehen, wie Menschen synthetische Stimmen wahrnehmen, kann die „Media Equation“ nach Reeves & Nass (1996) genutzt werden. Sie sagt aus, dass Menschen automatisch soziale Skripte auf die Interaktion mit Medientechnologien übertragen, wie beispielsweise einen sprechenden Computer. Soziale Skripte bezeichnen mentale Modelle für die Interaktion mit anderen (Honeycutt & Bryan, 2010; Schank & Abelson, 1977). Wenn der Computer spricht, entstehe beim Rezipienten automatisch eine soziale Situation, die denselben psychologischen Mechanismen folge, die für menschliche Interaktion evolutionär ausgebildet wurden (Reeves & Nass, 1996). Nass & Moon (2000) beschreiben diese automatische Aktivierung sozialer Skripte als „Mindlessness“ (Gedankenlosigkeit) und unterscheiden sie explizit

von bewusstem Anthropomorphismus, der Übertragung von menschlichen Eigenschaften auf nicht-menschliche Entitäten. Folglich ist „Mindlessness“ ein unbewusster, automatischer Prozess, durch den soziale Reize ausgelöst werden. Gambino et al. (2020) argumentieren, dass sich die Mensch-Medien-Kommunikation seitdem weiterentwickelt hat und Menschen nicht gedankenlos soziale Skripte der Mensch-zu-Mensch-Interaktion auf Medien anwenden. Vielmehr haben Menschen eigene soziale Skripte für die Mensch-Medien-Kommunikation entwickelt. Für die vorliegende Untersuchung erscheint die Position von Gambino et al. (2020) relevanter, da die Kommentare darauf hindeuten, dass Zuschauer auf YouTube eigene Skripte für die Interaktion mit synthetischen Stimmen entwickelt haben.

Laut Sundar & Nass (2000) behandeln Nutzer nicht den Programmierer hinter einem Computer, sondern den Computer selbst als Informationsquelle. Übertragen auf synthetische Stimmen auf YouTube könnte das bedeuten, dass Zuschauer sich nicht dem Menschen hinter dem Kanal, sondern der Stimme selbst als kommunikative Quelle orientieren.

Die Zuschreibung der Stimme zu einer wahrgenommenen Quelle hat außerdem einen signifikanten Einfluss darauf, wie eine Stimme bewertet wird (Stern et al., 2006). Wenn die Quelle der Stimme (Mensch oder Roboter) mit dem Stimmtyp (menschlich oder robotisch) konsequent war, wurden Botschaft und Stimme besser bewertet. So kann in Teilen erklärt werden, wie eine KI-Stimme in einer Funktion akzeptabel und in einer anderen irritierend wirken kann.

Auch die Kennzeichnung, dass eine Stimme KI-generiert ist und der damit einhergehende Label-Effekt zeigen, dass soziale Reaktionen unterschiedlich ausfallen (Fan & Liu, 2025, S. 6). Die Studie zeigt auch, dass es relevant für die Analyse von Kommentaren ist, ob bei gegebenen Videos die KI-Generierung der Stimme gekennzeichnet ist oder nicht. So werden gekennzeichnete Stimmen als weniger authentisch und vertrauenswürdig bewertet und der „Media Equation“-Effekt abgeschwächt.

Parasoziale Interaktion beschreibt die Illusion eines direkten und wechselseitigen Austauschs während der Mediennutzung, bei der der Zuschauer mit der Medienfigur interagiert, als stünden sie sich gegenüber. Eine parasoziale Beziehung ist die durch wiederholte parasoziale Interaktion gebildete, langfristige emotionale Bindung an diese Figur (Horton & Wohl, 1956). Auf Hartmann und Goldhoorn (2011) zeigten experimentell, dass parasoziale Interaktion auch in asynchronen Medienkontexten wie Videos messbar ist. Stimmklone können für Fans keine vollständig immersive Erfahrung erzeugen, aber parasoziale Beziehungen weiter vorantreiben (Kang et al., 2025). Es wurde kollektiv ausgehandelt, ob die Stimme als eine authentische Repräsentation der Medienfigur akzeptiert wurde. Diese Aushandlung lässt sich auf YouTube-Videos und Kanäle übertragen, wenn die Stimme in der Funktion der Persona ist. Eine Ausnahme könnten Kanäle bilden, die nicht auf einer echten Stimme und Person basieren, wodurch die KI-Stimme automatisch als Repräsentation der Medienfigur akzeptiert wird.

2.3 Funktion der Stimme im Nutzungskontext

Unabhängig vom Inhalt übertragen menschliche Stimmen soziale Signale, deshalb kann eine „neutrale“ Stimme nicht alle sozialen Funktionen gleich gut erfüllen (Cambre & Kulkarni, 2019, S. 2). Daraus ist zu schließen, dass auch synthetische, menschenähnliche Stimmen in verschiedenen Funktionen unterschiedlich gut erfüllen. Herauszufinden ist hierbei auch, ob synthetische Stimmen in bestimmten Funktionen konstant anders bewertet werden, als solche von menschlichen Sprechern, selbst wenn die Art der Stimme (zum Beispiel Tonlage, Geschlecht & Geschwindigkeit) gleich sind.

Dass Stimme in Medien verschiedene Formen annehmen kann, beschrieb Chion (1999) mit den drei Grundweisen der filmischen Stimme. Diese sind die Figurenstimme, bei der die gehörte Stimme zur Figur im Bild gehört, die Voice-Over-Stimme, die zu einer sprechenden Figur aus der Filmwelt gehört, die gerade nicht im Bild ist und die Voice-off-Stimme, die oft die Erzählerstimme ist. Diese Unterscheidung hilft aufzuzeigen, dass die Bedeutung einer Stimme auch aus ihrer Beziehung zu Bild entstehen kann. Weitergedacht und auf YouTube übertragen hängt die Wahrnehmung der Stimme in YouTube-Videos davon ab, welche Funktion und folglich Beziehung sie zu dem Gesagten hat.

2.3.1 Erwartungen an Authentizität in den Funktionen

In dieser Arbeit werden drei Funktionen einer Stimme in YouTube-Videos definiert. Die unterschiedlichen Stimmfunktionen haben Auswirkungen darauf, welche Erwartungen an Authentizität Zuschauer haben. Kozloff (1988) argumentiert, dass auch die Erzählstimme eine Position einnimmt, in der sie entscheidet, was das Publikum weiß und was nicht. Bezogen auf YouTube kann man daraus schließen, dass auch hier eine Stimme rein in der Funktion des Erzählers Authentizitätserwartungen erzeugt. Diese beziehen sich anders als bei der Stimme als Persona eher auf die Glaubwürdigkeit der Informationen. Nach Riboni (2020, S. 133) gibt es einen gemeinsamen Kern von Authentizitätsmerkmalen auf YouTube, auch über verschiedene Genres hinweg. Entgegen der Annahme, dass Authentizität an Nicht-Kommerzialität gebunden sei, zeigt sich, dass kommerzielle Interessen in die Identität integriert werden können, ohne dass dies als nicht authentisch wahrgenommen wird (Riboni, 2020, S. 135–136). Ein ähnlicher Effekt zeigt sich in der auditiven Wahrnehmung von KI: Wenn die Stimme als rein funktionales Werkzeug dient, kann eine maschinell-synthetische Stimme unter Umständen sogar positiver und passender wahrgenommen werden als eine menschliche (Im et al., 2023, S. 1, 5).

2.3.2 Einteilung von Stimmen nach Funktion

Ursprünglich sollten Videos zur Analyse nach Genre kategorisiert werden. Zwar spielen bestehende Kategorien und Formate inhaltlich eine Rolle, aber können nicht als alleiniger Bezugsrahmen dienen, da sie in der Praxis zu starke Überschneidungen aufweisen (Jost, 2024,

S. 89, 95). Die Einteilung nach Funktion der Stimme im Video (Nutzungskontext) ist hingegen klarer trennbar. Dass die übergeordnete Funktion und situativer Kontext einer Interaktion eine größere Relevanz als das Genre haben, zeigt auch das Modell zur ‚Vocal Person' von Noufi et al. (2025, S. 7, 9). Besonders für die Stimme als Persona ist der Label-Effekt (Fan & Liu, 2025) zusätzlich relevant, da durch gezieltes Kennzeichnen der Stimme als Figur, die Wahrnehmung verändert werden kann. Dieser Effekt kann beim Vergleich von Genres nicht so klar gemacht werden. Auch die Glaubwürdigkeit der Stimme ist von der Funktion und dem Nutzungskontext abhängig (Noufi et al., 2025, S. 9). Während Videos geschaut werden, ist die Funktion der Stimme durchgehend präsent, hingegen ist das Genre oft nur nebensächlich.

2.4 Kommentare als performative Rezeptionsdaten

Kommentare von Nutzern in sozialen Medien stellen eine wertvolle Informationsquelle über die Ansichten, Stimmungen, Bewertungen und Erfahrungen der Nutzer dar (Thelwall, 2018). Dennoch muss beachtet werden, dass nur etwa 78 bis 83 % der Kommentare für eine Analyse (zum Beispiel der Rezeption) relevant sind (Möller et al., 2024, S. 178). Bei den irrelevanten Kommentaren handelt es sich unter anderem um Spam-Kommentare (Poché et al., 2017). Um bei einer Sentimentanalyse kein verzerrtes Bild zu bekommen, muss aus diesem Grund die Relevanz gegebener Kommentare festgestellt werden. Kommentare ermöglichen Einblicke in Diskurs und Sentiment, können aber keine repräsentativen Aussagen über die Gesamtheit der Zuschauer machen (Thelwall, 2018). Anders als standardisierte Umfragedaten sind YouTube-Kommentare daher exploratives, qualitativ nutzbares Material, welches methodisch bewusste Sampling- und Analysestrategien erfordert. Auch ist zu beachten, dass Daten wie Nutzerkommentare auf YouTube möglicherweise keine Aussagekraft außerhalb der Plattform haben (Sui et al., 2022, S. 9–10).

Aussagen in YouTube-Kommentaren sind keine einfache Meinungsäußerung, sondern in einen Partizipationsrahmen eingebettet; sie richten sich also nicht nur an den Videoersteller, sondern auch an andere Kommentierende und stille Leser (Dynel, 2014, S. 46–47). Marwick & Boyd (2011) beschreiben Äußerungen in sozialen Medien als an ein ‚imaginiertes Publikum' gerichtet, das formt, was und wie geschrieben wird. Nutzerkommentare finden in einem sozialen Kontext statt und sind eine Art der sozialen Interaktion. Soziale Interaktion ist immer performativ, Menschen steuern also aktiv den Eindruck, den sie bei anderen hinterlassen wollen (Goffman, 1956). Bullingham & Vasconcelos (2013) übertragen dieses Modell auf digitale Plattformen und bestätigen, dass Menschen auch hier bewusst ihre präsentierte Identität wählen. Hogan (2010) argumentiert, Kommentare seien keine synchrone Situation, sondern asynchrone Ausstellungen. Anders als spontane Äußerungen sind Kommentare also Artefakte, die bestehen und einsehbar bleiben.

Nutzerkommentare auf YouTube finden in der Öffentlichkeit statt. Um zu verstehen, wie öffentliche Meinung sich bildet und warum sie möglicherweise nicht die Meinung der Mehrheit widerspiegelt, dient die Schweigespirale nach Noelle-Neumann (1974). Ihr zufolge schweigt, wer glaubt, in der Minderheit zu sein und wer sich durch die wahrgenommene Mehrheitsmeinung bestärkt fühlt, äußert sich. Auf YouTube kann dieses Phänomen neben der Anzahl an Kommentaren auch durch die Funktion Kommentare zu liken und die damit einhergehende „Top-Kommentare"-Funktion verstärkt werden. Allerdings kann so möglicherweise durch gezieltes Löschen von Kommentaren durch den Videoersteller auch die öffentliche Meinung in den Kommentaren gelenkt werden. Selbst bei objektiv gleichwertigem Inhalt wird die Wahrnehmung eines Videos durch soziale Informationen beeinflusst, so erhöhen viele Likes und positive Kommentare beispielsweise den Genuss an Videos (Möller et al., 2021, S. 42). Auch der Videokontext hat Auswirkung auf die Art der Interaktion in Kommentaren. Unterhaltungsvideos erhalten nach Möller et al. (2019, S. 523) neutralere, aber insgesamt mehr Kommentare im Verhältnis zu den Aufrufzahlen, während Zuschauer bei politischen Videos mit höherer Valenz reagieren. Laut einer Umfrage von Khan (2017, S. 243) lesen Nutzer Kommentare vornehmlich zur Informationsbeschaffung und schauen Videos zur Unterhaltung. Auch zeigte sie, dass das Geschlecht einen Einfluss auf die Interaktion mit Videos haben kann.

Die Interpretation von YouTube-Kommentaren sollte immer mit Vorsicht stattfinden. Nutzerkommentare sind valide Daten für die öffentliche Aushandlung von Wahrnehmungsurteilen einer investierten Gruppe, nicht aber für individuelle Einstellung oder Repräsentativitätsansprüche (Kozinets, 2015; Thelwall, 2018).

2.5 Ableitung für die Arbeit

Aus dem bisherigen Forschungsstand lässt sich für diese Arbeit die Vorannahme ableiten, dass die Wahrnehmung synthetischer Stimmen, je nach Funktion in Videos auf YouTube, variiert und Bewertungsdimensionen unterschiedlich gewichtet werden.

Die Funktionen der Stimme bilden unterschiedliche Kontexte. Durch sie wird die Erwartung des Publikums angepasst und es ergeben sich unterschiedliche Bewertungsmaßstäbe. Diese Maßstäbe und somit unterschiedliche Wahrnehmung und Akzeptanz zeigen sich in den Kommentaren unter relevanten YouTube-Videos. Auch wenn die Kommentare keine Grundlage zur Analyse sind, was die Mehrheit vertritt, können sie als Indikatoren für Stimmung engagierter Zuschauer gegenüber synthetischen Stimmen in den Videos wahrgenommen werden.

Die Untersuchung geht auf Basis bestehender Forschung mit der Annahme ins Feld, dass für die Stimme als Werkzeug primär die funktionale Stimmigkeit, Verständlichkeit und Effizienz relevant sind, während die personale Authentizität wenig Notwendigkeit aufweist. Bei der Stimme als Erzähler wird voraussichtlich der Fokus auf Glaubwürdigkeit, Stimmigkeit,

Ausdruck und der narrativen Angemessenheit liegen. Die Stimme als Persona muss hingegen möglicherweise starken Authentizitätserwartungen entsprechen, konsistente Identität zeigen und transparent auftreten.

3 Methodik

3.1 Forschungsdesign und Erkenntnisinteresse

Da es bisher keine Forschung zur Rezeption synthetischer Stimmen nach Funktion in YouTube-Videos durch Analyse von Kommentaren gibt, ist ein exploratives Design angemessen, um Hypothesen zu generieren. Explorative Forschung ist nach Stebbins (2001) ein vorab geplantes Unterfangen, das durch offene und flexible Grundhaltung darauf abzielt, neue Erkenntnisse zum tieferen Verständnis zu entdecken. Die Untersuchung hat demnach nicht den Anspruch, klare Hypothesen zu beantworten, vielmehr sollen erste Hypothesen aus den Ergebnissen generiert werden und Muster, Ideen und Zusammenhänge gefunden werden. Um mögliche Hypothesen zu finden, wie und warum Stimmen in verschiedenen Funktionen wahrgenommen werden und wie Medienersteller dies nutzen können, wird eine inhaltlich strukturierende qualitative Inhaltsanalyse der Kommentare nach Kuckartz (2016) durchgeführt. Dabei werden zuerst Hauptkategorien aus bestehender Forschung (deduktiv) gebildet und anschließend aus den Daten (induktiv) relevante Unterkategorien. Um Unterschiede zwischen den Funktionen zu finden, werden sie systematisch verglichen. Ohne einen Vergleich bliebe unklar, ob Reaktionen funktionsspezifisch sind. YouTube-Kommentare sind kontextabhängige, performative Äußerungen (Bullingham & Vasconcelos, 2013) deren Inhalt qualitativ analysiert werden muss, um nuancierten Aufschluss über mögliche Hypothesen zu geben. Der vergleichende Aufbau folgt direkt der Fragestellung, um Aussagen über Funktionsabhängigkeit zu machen.

3.2 Herleitung und Definition funktionaler Stimmkategorien

3.2.1 Funktionskategorienbildung

Die drei Stimmfunktionen werden theoriegeleitet aus bestehender Literatur hergeleitet. Als Ausgangspunkt dient Chions (1999) Unterscheidung von Stimmpositionen im audiovisuellen Medium, insbesondere seine Differenzierung zwischen Ton, der innerhalb der erzählten Welt eines Films existiert, wie die Stimme einer Figur und dem, der nur vom Publikum hörbar ist, zum Beispiel ein Erzähler. Daraus lassen sich für diese Untersuchungen zwei Stimmfunktionen herleiten: die Erzählstimme (Chion, 1999; Kozloff, 1988) und die Figurenstimme beziehungsweise Persona (Chion, 1999; Noufi et al., 2025). Die dritte Stimmfunktion, die Stimme als Werkzeug, spielt in klassischen Filmen keine Rolle und ist deshalb nicht aus filmtheoretischen Ansätzen herleitbar. Für YouTube-Inhalte ist sie jedoch analytisch sinnvoll. Ihre Grundlage bilden in dieser Untersuchung Cambre & Kulkarni (2019), die zeigen, dass Stimmen soziale Signale übertragen, woraus folgt, dass Kontexte ohne soziale Anforderungen eine eigene funktionale Kategorie darstellen. Bestätigt wird das durch Im et al. (2023), die zeigen, dass

synthetische Stimmen in rein funktionalen Aufgaben eigenständig und in einigen Fällen sogar besser als die Stimmen von Menschen bewertet werden.

Im Folgenden werden die Definitionen der verschiedenen Stimmfunktionen operationalisiert, um sie so in Videos eindeutig zuzuordnen. Dabei müssen nicht alle Punkte der Definitionen vollständig zutreffen, um einer Stimme eine Definition zuzuordnen. In Fällen, bei denen zwei Definitionen zutreffen können, wird eine Hauptfunktion identifiziert. Die Hauptfunktion ist immer die, von der ein größerer Teil der Definition einer Funktion zutrifft.

3.2.2 Stimme als Werkzeug

Die Stimme in der Funktion als Werkzeug erfüllt eine rein funktionale Aufgabe, bei der sie Informationen transportiert, ohne eine Beziehung zwischen Zuschauer und Sprecher herzustellen. Die Inhalte sind oft auch ohne die Stimme zugänglich, durch Text im Bild, Anleitungen, Listen oder Definitionen. Die Stimme ist austauschbar, anonym und der Sprecher hat keine wahrnehmbare Identität oder diese ist nicht relevant. Im Video wird keine Verbindung zwischen Inhalten und Stimme kommuniziert. Die Stimme ist ein Ausgabemittel und nicht der Autor. Die Stimme wechselt innerhalb eines Kanals möglicherweise. Sie wird keinem Avatar, Charakter oder Ähnlichem zugeordnet. Die Stimme bewertet und kommentiert nicht.

Typische Formate sind Anleitungen (Tutorials) und automatische Erklärungen verschiedener Themen.

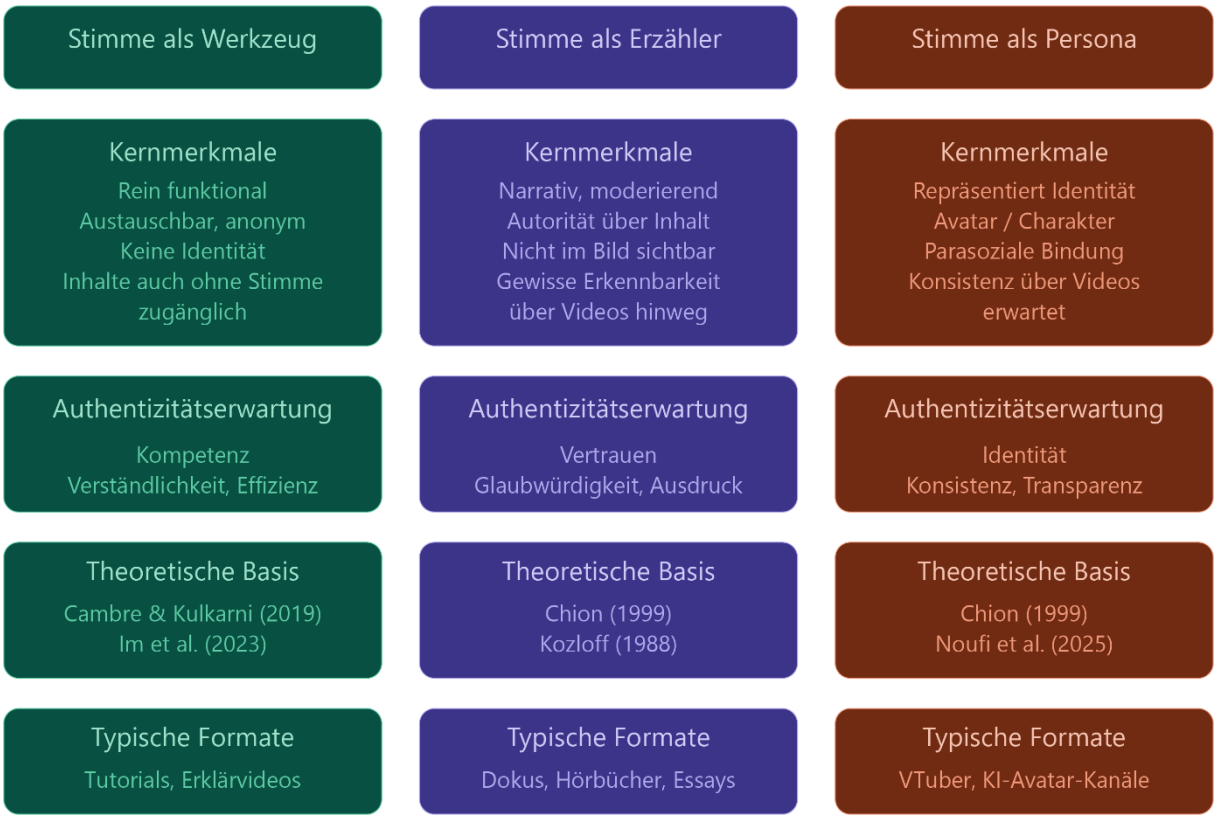
3.2.3 Stimme als Erzählstimme

In der Funktion des Erzählers nimmt die Stimme eine narrative und moderierende Position ein. Sie führt durch Inhalte und erklärt, bewertet und rahmt diese ein. Funktional hat sie eine Autorität über den Inhalt, aber keine Identität, die über das Video hinausgeht. Die Stimme übernimmt eine klassische Voice-Over- oder Off-Funktion (Chion, 1999). Sie ist nicht im Bild zu sehen, aber nimmt Einfluss darauf, was im Bild ist und wie es dargestellt wird. Ein Kanal mit primär dieser Stimmfunktion in den Videos kann eine erkennbare redaktionelle Linie haben, ohne als Persönlichkeit aufzutreten. Die Stimme weist eine gewisse Erkennbarkeit über Videos eines Kanals hinweg auf, tritt aber nicht als Figur oder Persönlichkeit auf. Sie wird nicht explizit mit einer Person verknüpft und liest nicht lediglich vor, sondern nimmt Einfluss auf die Darstellung.

Typische Formate sind Dokumentationen, Hörbücher, Video-Essays, Nachrichten und Filmzusammenfassungen.

Abbildung 1 fasst die drei Stimmfunktionen mit ihren Kernmerkmalen, theoretischen Grundlagen und den erwarteten Authentizitätsdimensionen zusammen.

Abbildung 1
Übersicht über die drei funktionalen Stimmkategorien



Anmerkung. Eigene Darstellung auf Basis von Chion (1999), Cambre und Kulkarni (2019), Im et al. (2023) und Noufi et al. (2025). Die Kategorien schließen sich nicht gegenseitig aus; bei Grenzfällen wird eine Hauptfunktion bestimmt (siehe Abschnitt 3.3.5).

3.2.4 Stimme als Persona

Die Stimme repräsentiert eine Identität und ist einem Avatar, Charakter oder Ähnlichem zugeordnet, der im Video zu sehen ist und spezifisch für den Kanal ist. Sie ist grundlegend für eine parasoziale Bindung zum Kanal. Eine solche Stimme hat explizit einen Namen und ist mit einer visuellen Identität verbunden. Sie ist an der Wiedererkennbarkeit des Kanals beteiligt und wird durch Zuschauer direkt als Figur oder Person in den Kommentaren angesprochen. Die Stimme ist über Videos hinweg gleichbleibend, da eine Konsistenz erwartet wird. Die Stimme beeinflusst durch ihre Identität die Inhalte.

Typischerweise wird die Stimme in dieser Funktion bei VTubern, Kanälen mit KI-Avataren oder animierten Inhalten genutzt.

3.3 Fallauswahl und Zuordnung der Videos

3.3.1 Identifikation von Videos

Videos wurden durch ein mehrstufiges Vorgehen identifiziert. Um Kanäle und Videos zu finden, gab es hierfür zwei Ansätze. Der erste Ansatz war die YouTube-Suche. Hier wurden zuvor Suchbegriffe definiert, die entweder direkt nach Angaben zur Nutzung von synthetischen

Stimmen suchten, darunter fallen zum Beispiel „ai voiceover“, „tts channel“ oder „voice over by ElevenLabs“, oder nach Videos gesucht, die inhaltlich mit KI-Stimmen zu tun haben. Dafür wurden Suchbegriffe wie „ai avatar“, „ai tutorial“ oder „tts-tuber“ genutzt. Da die Auswahl an Kanälen, die durch diese Suche zustande kam, jedoch begrenzt war und nicht den weiteren YouTube-Kontext, sondern hauptsächlich Inhalte, die sich direkt mit künstlicher Intelligenz befassen, abbildete, wurden weitere Ansätze zur Diversifizierung der Kanäle benötigt. Dafür wurde Ansatz zwei durchgeführt. Bei diesem wurden externe Quellen als Einstieg genutzt. Reddit-Threads, Artikel über KI-YouTube-Kanäle, Twitter/ X-Diskussionen über synthetische Stimmen auf YouTube verweisen oft direkt auf konkrete Beispiele. Durch diesen Ansatz wurden bekannte Kanäle identifiziert, die sich nicht mit dem Thema künstliche Intelligenz befassen. Insgesamt wurden zu Beginn also die ersten Kanäle identifiziert. Um weitere Kanäle zu finden, wurden drei dieser Kanäle als Startpunkt genutzt, je einer dieser Kanäle stand für eine der drei Stimmfunktionen. Von dem Kanal aus wurde eine algorithmisch geleitete Suche durchgeführt. Dafür wurde in einem Inkognito-Fenster, ohne auf YouTube eingeloggt zu sein, entweder auf ein Video geklickt und von da aus auf empfohlene Videos, um so ähnliche Inhalte zu finden oder wenn möglich auf der Kanalseite auf empfohlene Kanäle weitergleitet. In beiden Fällen wurden im Anschluss die Videos auf KI-Stimme analysiert und dem Korpus hinzugefügt, falls eine KI-Stimme bei einem der so gefundenen Videos und Kanäle identifiziert werden konnte. In einigen Fällen umschloss die Suche auch die Verlinkungen in Beschreibungen von Videos mit KI-Stimme und die Kommentare. Beispielsweise wenn ein Video eine weitere Stimme oder Person umfasst, die einen separaten Kanal hat. Wenn bei einem Kanal oder Video keine KI-Stimme identifiziert werden konnte, wurde der Schritt rückgängig gemacht und vom vorigen Punkt weitergesucht. Im Anschluss wurde dieser Schritt weiter durchgeführt, bis eine Auswahl von etwa zehn Kanälen pro Stimmfunktion entstand. Danach wurde eine Auswahl von zwei bis drei Videos pro Kanal erstellt, die diesen Kanal bestmöglich abdecken. Um das zu erreichen, wurden die meistgeklickten Videos, die relevant für die Untersuchung sind und alle Kriterien (wie im kommenden Unterpunkt definiert) erfüllen, genutzt. So wurden beispielsweise Videos entfernt, die keine KI-Stimme nutzen oder nicht dem Durchschnitt der Videos auf gegebenem Kanal entsprechen, zum Beispiel Musikvideos auf einem Gaming-Kanal.

Für alle Videos und Kommentare wurden Datum und Uhrzeit dokumentiert, an dem sie zuletzt aufgerufen wurden, da sich die Informationen ständig ändern können durch Löschung von Kommentaren, Videos oder ganzen Kanälen.

3.3.2 Identifikation von KI-Stimmen

Da die Nutzung von KI-Stimmen in vielen Fällen nicht explizit angegeben ist und dies zu anderer Wahrnehmung bei Zuschauern führen kann, ist das Teil der Analyse dieser

Untersuchung. Um also auch diese Fälle abzudecken, werden auch sie identifiziert und gekennzeichnet.

Die Identifikation erfolgt in Abstufungen. Der erste und stärkste Nachweis ist die Angabe durch den Videoersteller. Diese kann erfolgen durch Erwähnung im Video (auditiv oder visuell), in einem Kommentar des Erstellers, auf der Kanalseite oder in der Beschreibung eines Videos. Bei diesen Angaben kann eindeutig davon ausgegangen werden, dass eine KI-Stimme genutzt wurde.

Eine etwas abgeschwächte Version ist die Kennzeichnung, dass KI in der Erstellung des Videos genutzt wurde, ohne konkrete Erwähnung der KI-Nutzung für die Stimme. Dies kann auch an allen zuvor genannten Punkten stattfinden. Außerdem gibt YouTube dem Ersteller die Möglichkeit, bei einem Video zu kennzeichnen, wenn große Teile davon künstlicher Intelligenz nutzen.

Eine Grauzone sind Videos und Kanäle, die ihre Nutzung von KI nicht angeben, deren Inhaber aber an anderer Stelle erwähnen, dass sie KI für die Stimme nutzen. Diese werden in der Analyse als Kanäle mit KI-Stimme aufgenommen, allerdings bei der Analyse der Zuschauerwahrnehmung als nicht gekennzeichnet markiert.

Die letzte Stufe, die bei der Identifikation durchlaufen wird, ist eine Mischung aus der Wahrnehmung des Untersuchenden unter Zuhilfenahme von externen Webseiten zur Identifikation von synthetischen Stimmen. Dabei wird die Tonspur eines Videos heruntergeladen und bei mehreren Detektoren hochgeladen, darunter AI Voice Detector (Undetectable AI, o. J.), AI Voice Detector (AIVoiceDetector.com, o. J.), AI Speech Classifier (ElevenLabs, o. J.), Find AI Voice (FindAIVoice.com, o. J.) und AI Voice Detector (TruthScan, o. J.), um eine Einschätzung zu bekommen, ob es sich um eine synthetische Stimme handelt. Da die Zuverlässigkeit solcher Detektoren variiert und nicht unabhängig validiert ist, wurden die Ergebnisse nur als ergänzender Hinweis und nicht als alleiniger Nachweis genutzt. Besonders die letzte Stufe birgt jedoch Unsicherheit und wird deshalb mit Vorbehalt eingeschlossen und in der Analyse betrachtet.

3.3.3 Anzahl der Fälle

Für jede Stimmfunktion sollen etwa zehn Kanäle identifiziert werden, von denen drei Videos ausgewählt werden, um in die Untersuchung aufgenommen zu werden.

Die Anzahl analysierter Videos orientiert sich nicht an einem Repräsentativitätsanspruch, sondern dem Prinzip der theoretischen Sättigung. Nach Stebbins (2001) ist bei explorativer Forschung entscheidend, ob neue Fälle noch neue Erkenntnisse, Muster und Kategorien hervorbringen. Sobald zusätzliche Videos innerhalb einer Funktionskategorie keine neuen Analysedimensionen mehr eröffnen, gilt die Kategorie als gesättigt. Zwei bis drei Videos pro Kanal

wurden als ausreichend eingeschätzt, da das Ziel der Arbeit die Generierung erster Hypothesen ist und nicht die Absicherung verallgemeinerbarer Aussagen. Es wurde bewusst Variation innerhalb der Kategorien angestrebt, um nicht nur gleichartige Fälle zu analysieren.

3.3.4 Ein- und Ausschlusskriterien

Das wichtigste Einschlusskriterium ist, dass die Stimme in einem Video nachweislich oder sehr wahrscheinlich synthetisch generiert wurde. Auch gilt eine Mindestanzahl von 50 Kommentaren pro Video, zum Zeitpunkt der Erhebung, um eine hinreichende Materialbasis für die qualitative Analyse zu gewährleisten. Der Schwellenwert wurde bewusst niedrig angesetzt, um nicht ausschließlich reichweitenstarke Kanäle einzubeziehen. Kleinere Kanäle können hinsichtlich der Kommentarkultur und des Diskursverhaltens von großen abweichen und erweitern so die Varianz des Materials. Ein höherer Schwellenwert hätte diese Fälle systematisch ausgeschlossen und die Fallauswahl möglichweise zugunsten etablierter, algorithmisch begünstigter Kanäle verzerrt. Die Analyse beschränkt sich auf englischsprachige Videos und Kommentare. Diese Einschränkung ist aus mehreren Gründen methodisch begründet. Englisch ist die dominante Sprache des für diese Arbeit relevanten YouTube-Diskurses über KI-generierte Inhalte, wodurch die Verfügbarkeit geeigneten Materials am größten ist. Darüber hinaus sichert die Einsprachigkeit die Vergleichbarkeit des Korpus, da Übersetzungen semantische Nuancen verzerren und unterschiedliche Sprachräume unterschiedliche Kommunikationskulturen mitbringen können, die eine funktionsübergreifende Analyse erschweren würden. Für die Analyse sind nur hinreichend menschenähnliche synthetische Stimmen relevant. Diese sind erst durch neuronale TTS-Verfahren der jüngsten Entwicklungsstufe erreicht worden (Nussbaum et al., 2025; Orynbay et al., 2024), weshalb ausschließlich Videos ab 2022 berücksichtigt werden. Auch muss die Kommentarfunktion der Videos aktiviert sein und die Hauptfunktion muss eindeutig zuordenbar sein.

Ausgeschlossen werden Videos, bei denen keine KI-Stimme festzustellen ist, weder durch Kennzeichnung noch auditive Erkennung. Videos mit deaktiviertem Kommentarbereich, Musikvideos und rein visuelle Formate ohne Sprechfunktion werden nicht einbezogen. Videos werden nicht aufgrund fehlender Kommentare, die Relevanz für die Analyse haben, ausgeschlossen, da dieses Fehlen auch eine Aussagekraft hat.

3.3.5 Grenzfälle

Ein Video ist ein Grenzfall, wenn ihm mehr als eine Stimmfunktion zugeordnet werden kann, weil die Stimme gleichzeitig Merkmale mehrerer Funktionen erfüllt. In der Dokumentation werden solche Grenzfälle markiert. Ein Erklär-Kanal, der die Stimme als Werkzeug nutzt, kann durch das Nutzen eines Avatars, dem er diese Stimme zuordnet, beispielsweise gleichzeitig die Stimmfunktionen Werkzeug und Persona haben. In diesen Fällen wird eine Hauptfunktion

bestimmt. Grenzfälle werden erst dann ausgeschlossen, wenn nicht eine der Funktionen dominiert und deshalb keine Hauptfunktion bestimmt werden kann. So wird methodisch aufgezeigt, wo Grenzen der Stimmfunktionen als Kategorien liegen.

3.3.6 Zuordnung der Stimmfunktionen

Die Zuordnung erfolgt vor der Kommentaranalyse, um die Kategorisierung nicht unbewusst durch Kommentareinhalte zu beeinflussen. So kann die Gefahr eines Zirkelschlusses verringert werden, beispielsweise könnten Kommentare, die viel über Authentizität sprechen, fälschlicherweise ein Video der Stimmfunktion Persona erscheinen lassen. Die Zuordnung basiert ausschließlich auf dem Video selbst, dem Kanalkontext und expliziten Metadaten. Es wird dabei untersucht, wie die Stimme im Verhältnis zum Bild und Inhalt des Videos auftritt, wie sie über Videos hinweg eingesetzt wird und welche möglichen Aussagen diesbezüglich in Kanal- und Videobeschreibung zu finden sind. Hingegen sind Genre, Kanalname, Thema oder Kommentarinhalte explizit nicht Grundlage der Zuordnung.

Bei der Zuordnung werden drei Schritte durchlaufen. Zuerst wird geprüft, ob die Merkmale der Stimme als Persona zutreffen und die Negativmerkmale nicht zutreffen. Im Anschluss wird das gleiche für die Stimme als Erzählstimme und die Stimme als Werkzeug durchgeführt. Wenn eine der Stimmfunktionen eindeutig zuordenbar ist, wird das Video dieser zugeordnet. Falls ein Video keiner Stimmfunktion zugeordnet werden kann, wird es aus der Untersuchung entfernt. Bei Grenzfällen wird geschaut, welche der Stimmfunktionen im Video die Hauptfunktion ist, durch Abgleichen mit den Merkmalen und Identifizieren, welche Funktion präsenter auftritt.

Für jedes Video wird in der Dokumentation ein kurzes Zuordnungsprotokoll angelegt, bei dem definiert wird, um welche Stimme es sich handelt und im Fall von Grenzfällen, welche die Unterfunktion ist. Nach der Zuordnung aller Videos werden Fälle innerhalb der Stimmfunktionen nochmal verglichen, um zu prüfen, ob die Videos tatsächlich untereinander ähnlicher sind als zu Videos anderer Stimmfunktionen.

3.4 Datenerhebung und Materialkorpus

Bei den gesammelten Daten handelt es sich um öffentlich zugängliche YouTube-Kommentare, welche das Analysematerial bilden. Dabei ist primär der geschriebene Text relevant und nicht Nutzernamen oder Likes. Für jedes Video wurde der Zeitpunkt der Erhebung dokumentiert. Sekundär sind Daten wie Titel, Kanalname, Veröffentlichungsdatum, Aufrufzahlen und Erhebungszeitpunkt. Diese Daten dienen nicht der Analyse, sondern lediglich der Dokumentation und Nachvollziehbarkeit der Untersuchung. Der Videoinhalt selbst wird abseits von der Zuordnung der Stimmfunktion nicht analysiert.

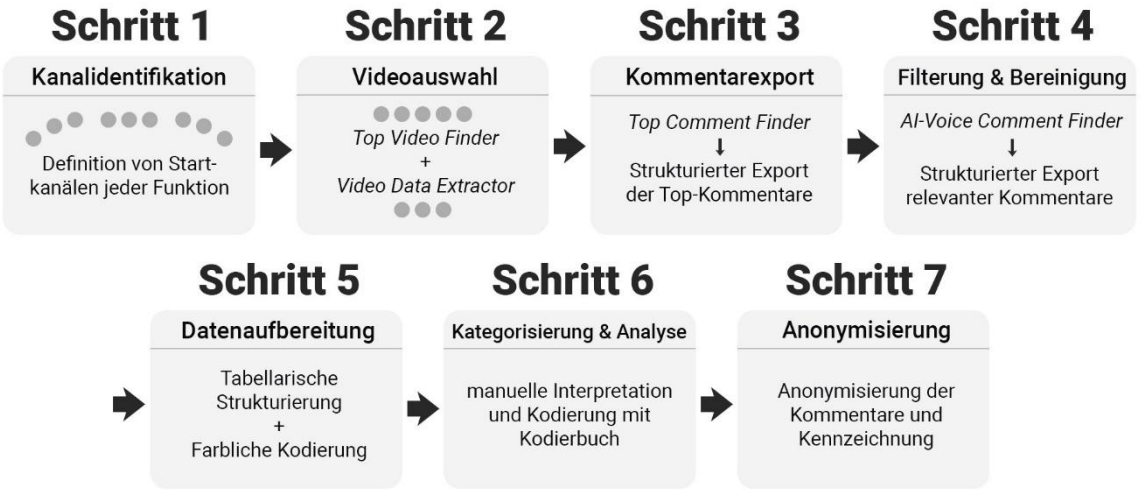
Pro Video wurden die zehn Top-Kommentare erfasst. Da Top-Kommentare auf YouTube nach Likes und Interaktionsrate angezeigt werden, repräsentieren sie den von der Community am stärksten bestätigten und für neue Zuschauer sichtbarsten Teil des Diskurses. Sie ermöglichen einen Vergleich mit dem Gesamtkorpus und können Hinweise auf Schweigespiraleffekte geben, wenn meistbestätigte Reaktionen systematisch von der Gesamttendenz abweichen (Noelle-Neumann, 1974). Antwortkommentare (Replies) wurden ausgeschlossen, da sie kontextuell vom jeweiligen Erstkommentar abhängig sind und ohne diesen oft nicht eigenständig interpretierbar sind. Ihr Einbezug würde zudem die Vergleichbarkeit zwischen Videos einschränken. Alle Kommentare, die sich inhaltlich mit der Stimme befassen oder künstliche Intelligenz kommentieren und auch auf die Stimme bezogen sein können, wurden durch gezielte Auswahl erfasst. Kommentare, die sich mit anderen Formen der KI in Videos befassen, wurden ebenso wie Spam-Kommentare herausgefiltert. Da es sich bei den Videos im Korpus ausschließlich um englischsprachige Videos handelt, wurden auch nur englischsprachige Kommentare berücksichtigt.

Der Korpus wurde tabellarisch dokumentiert. Für jedes Video wurden festgehalten: Videotitel, Video-URL, Veröffentlichungsdatum, Aufrufzahl, Stimmfunktion, KI angegeben (Ja/ Nein) und KI-Nachweis. Die erhobenen Daten wurden als Rohdaten gesichert und vor der Kodierung nicht verändert.

3.4.1 Durchführung der Datenerhebung

Abbildung 2

Ablauf der Datenerhebung und -aufbereitung



Die Kanalidentifikation erfolgte auf Basis des in 3.3 beschriebenen Vorgehens. Wie in Abbildung 2 dargestellt, dienten folgende Kanäle in Schritt 1 als Startpunkt: Neuro-Sama, Zentreya und Yellow Dude für Persona, Industrial Decay, Gates of Imagination und Grand Manors für Erzählstimme sowie Zolven, Money Groot und Everything Professor für Werkzeug. Die Suche über Kanalseiten-Empfehlungen lieferte keine Ergebnisse, weshalb die weitere

Kanalidentifikation ausschließlich über vorgeschlagene Videos und die YouTube-Suchfunktion erfolgte. Nachdem durch diese Suche ausreichend Kanäle identifiziert worden waren, wurden die Daten gesammelt.

Um Daten nicht händisch zu sammeln, wurden mehrere Apps mithilfe von Google AI Studio gebaut, die über die YouTube API Zugriff auf alle nötigen Daten hatten. Die erste Web-App ist Top Video Finder (siehe Anhang B), bei der Links zu den gesammelten Kanälen eingefügt werden und eine TSV-Datei der fünf meistgeklickten Videos der Kanäle, optional mit den jeweiligen Top-Kommentaren, ausgegeben wird. Durch diese Apps wurden für die Kanäle die relevanten Videos gesammelt und je drei Videos dem Datensatz hinzugefügt. Eine zweite Web-App namens Top Comment Finder (siehe Anhang B) wurde im Anschluss zum strukturierten Export der Top-Kommentare inklusive aller relevanter Daten als CSV-Datei genutzt. Für Fälle, in denen die fünf meistgeklickten Videos nicht drei relevante Videos enthielten, wurde die App Video Data Extractor (siehe Anhang B) genutzt, bei der Video-Links eingefügt werden und dann alle relevanten Daten als CSV-Datei ausgegeben werden. Als Letztes wurde die App AI-Voice Comment Finder (siehe Anhang B) verwendet, um für die gesammelten Videos alle Kommentare herauszufiltern, die entweder „ai“ oder „voice“ in irgendeiner Form beinhalteten. Im Anschluss wurden alle so erhobenen Kommentare entfernt, die mit „ai“ im Kommentar explizit nur auf andere Formen der KI-Nutzung, außerhalb des Untersuchungsinteresses, Bezug nahmen. Zur ordentlichen Sammlung der Daten wurden sie in Excel eingefügt und mithilfe des Claude Add-Ins (Anthropic, 2025) visuell aufbereitet, indem die Stimmfunktionen farblich kodiert und die Daten tabellarisch strukturiert wurden. Für die Kategorienzuordnung wurde eine separate Excel-Datei genutzt. Die Interpretation und Kodierung fanden im Anschluss ausschließlich manuell durch den Verfasser statt. KI-gestützte Tools wurden lediglich für die technische Datenerhebung und tabellarische Aufbereitung genutzt. Alle Kommentare wurden im Anschluss anonymisiert.

3.5 Qualitative Inhaltsanalyse und Kodierlogik

Zur Analyse der YouTube-Kommentare wurde eine inhaltlich strukturierende qualitative Inhaltsanalyse nach Kuckartz (2016) genutzt. Für die Fragestellung eignet sich dieses Verfahren aus mehreren Gründen. YouTube-Kommentare sind kontextabhängige, performative Äußerungen, deren Bedeutung nicht durch Zahlen oder Sentimentanalyse erfasst werden kann. Nuancen, die für den Vergleich von Funktionskontexten relevant sind, brauchen eine qualitative Interpretation des Materials. Ein Kommentar kann beispielsweise auf den Klang oder die Tatsache, dass KI eingesetzt wurde, reagieren, was nur durch diese Art der Analyse klar erfasst wird. Die strukturierende Inhaltsanalyse bietet außerdem einen systematischen Rahmen, durch den die Vergleichbarkeit zwischen den Stimmfunktionen gewährleistet und nachvollziehbar gemacht werden kann.

Die Kategorienbildung erfolgte kombiniert deduktiv-induktiv. Es wurden zuerst aus der bestehenden Forschung die relevantesten Bewertungsdimensionen synthetischer Stimmen herausgearbeitet (deduktiv). Daraus wurden die Hauptkategorien der Analyse abgeleitet. Da diese Kategorien, die aus Laborstudien entwickelt wurden, jedoch nicht direkt auf YouTube-Kommentare übertragen werden können, da es sich nicht um standardisierte Bewertungen, sondern öffentliche Äußerungen handelt, wurde auch induktiv gearbeitet, also aus den Daten heraus neue Unterkategorien gebildet. So soll sichergestellt werden, dass die Kategorien für die erhobenen Daten relevant sind und nicht durch Hauptkategorien abgedeckte Reaktionsmuster abgebildet werden.

Es wurden vier relevante Hauptkategorien deduktiv gebildet. Die erste Kategorie ist Klang und Natürlichkeit. Sie erfasst Kommentare, die sich explizit mit der Klangqualität der Stimme beziehen. Das können wahrgenommene Roboterhaftigkeit, fehlender emotionaler Ausdruck oder auch angenehmer Klang sein. Natürlichkeit gilt laut Hinterleitner (2017) als stärkster Prädiktor für die Gesamtbewertung synthetischer Stimmen. Die zweite Kategorie ist Glaubwürdigkeit und Vertrauen. Sie erfasst Kommentare, die Zweifel an der Vertrauenswürdigkeit des Inhalts oder der Quelle beschreiben. Nachweislich ist die Glaubwürdigkeit synthetischer Stimmen kontextabhängig (Gong, 2023; Maltezou-Papastylianou et al., 2025a), weshalb diese Kategorie besonders relevant für den Vergleich der unterschiedlichen Stimmfunktionen ist, gerade bei der Stimme als Werkzeug und Erzählstimme, da hier die Qualität der Informationen im Vordergrund steht. Die dritte Kategorie ist Authentizität und Identität. Sie unterscheidet sich von Glaubwürdigkeit, da sie sich nicht mit dem Inhalt, sondern der Person hinter der Stimme befasst. Kommentare könnten zum Beispiel das Vermissen einer echten menschlichen Präsenz beschreiben oder die Passung zwischen Kanalidentität und Stimme bewerten. Besonders für die Persona-Funktion ist diese Kategorie zentral (Kang et al., 2025; Riboni, 2020). Die vierte und letzte Hauptkategorie ist Reaktion auf KI-Einsatz. Sie erfasst Kommentare, die nicht auf den Klang der Stimme, sondern den KI-Einsatz als solchen reagieren. Grundlage dieser Kategorie ist der Label-Effekt nach Fan & Liu (2025). Folglich kann das Wissen über die Nutzung von einer KI-Stimme die Rezeption unabhängig von klanglicher Qualität beeinflussen. Weil ein erheblicher Teil der Kommentare im Material keiner der drei ersten Kategorien zugeordnet werden kann, ist diese Kategorie notwendig, um die Rezeption als Ganzes ohne Informationsverlust darzustellen.

Es wurde auf weitere Kategorien, die in der Forschung vermehrt auftauchen, wie Verständlichkeit, Sympathie oder Uncanny Valley als eigenständige Hauptkategorien verzichtet. Hinterleitner (2017) beschreibt, dass die Verständlichkeit in den letzten Jahren deutlich verbessert wurde, allerdings immer noch eine wichtige Bewertungsdimension bleibt. Sie fließt bei dieser Arbeit in die Kategorie Natürlichkeit und Klang ein. Uncanny Valley und parasoziale Reaktionen wurden als mögliche induktive Unterkategorien behandelt, falls sie im Material auftreten.

Kodiert wurde anhand eines Kodierbuchs, das für jede Hauptkategorie eine Definition, Ankerbeispiele und Abgrenzungsregeln zu anderen Kategorien enthält. Jeder Kommentar wurde zunächst darauf geprüft, ob er für die Analyse relevant ist, wie zuvor beschrieben. Anschließend wurden relevanten Kommentaren einer oder mehreren Hauptkategorien zugeordnet, da sie häufig mehrere Dimensionen ansprechen. Zusätzlich wurde die Valenz festgehalten: positiv, negativ oder ambivalent. So sollen funktionsübergreifende Vergleiche ermöglicht werden, beispielsweise ob Natürlichkeit bei der Stimme als Erzählstimme positiver bewertet wird als bei der Stimme als Werkzeug.

Mehrdeutige Kommentare wurden durch eine Entscheidungsregel behandelt. Wenn ein Kommentar mehreren Kategorien zugeordnet werden kann, wird er allen zugeordnet. Die Kodierung erfolgte in drei Durchläufen. In der Erstkodierung wurde zunächst eine Auswahl von Kommentaren aus allen drei Stimmfunktionen kodiert, um das Kodierbuch zu erproben und auf das Material anzupassen. Im Anschluss wurden die Kategoriendefinitionen geschärft, Ankerbeispiele ergänzt und Abgrenzungsregeln präzisiert. Dabei entstanden auch erste induktive Unterkategorien, die im Material aufgetaucht waren, aber durch die deduktiven Hauptkategorien nicht abgedeckt wurden. In der anschließenden Feinkodierung wurde der gesamte Korpus vollständig kodiert. Zweifelsfälle wurden angemerkt. Das vollständige Kodierbuch findet sich im Anhang A.

3.6 Reflexion der Methode und ihrer Grenzen

3.6.1 Repräsentativität von Kommentaren

Die Mehrheit der Zuschauer kommentiert nicht, Kommentare stammen meist nur von einer kleinen aktiven Teilgruppe (Thelwall, 2018), sie können also nicht als repräsentativ für alle Zuschauer gesehen werden. Kommentierende unterscheiden sich von passiven Zuschauern. Sie sind stärker involviert, haben möglicherweise stärkere Meinungen und mehr Vorwissen über KI. Auch bieten Kommentare meist keine Aussagen über demografische Variablen wie Alter, Geschlecht oder Vorwissen, welche die Wahrnehmung synthetischer Stimmen beeinflussen können (Bruder et al., 2025). Die Untersuchung macht keine Repräsentativitätsansprüche und behandelt Kommentare als Indikatoren für die Stimmung, nicht als Abbild der Gesamtrezeption (Thelwall, 2018).

3.6.2 Performativität von Kommentaren

Kommentare sind keine spontanen, privaten Äußerungen, sondern öffentliche Artefakte (Hogan, 2010). Sie richten sich nicht nur an den Ersteller, sondern auch an andere Lesende und somit ein vorgestelltes Publikum (Marwick & Boyd, 2011). Menschen versuchen aktiv den Eindruck zu steuern, den sie bei anderen hinterlassen, auch in digitalen Kontexten (Bullingham &

Vasconcelos, 2013). Das bedeutet, dass Kommentare zugespitzt, übertrieben negativ oder positiv ausfallen können, um Zustimmung zu erhalten oder einer Position mehr Ausdruck zu geben. Besonders negative Kommentare über KI-Stimmen könnten durch soziale Umstände verstärkt werden, wenn KI-Kritik in einer Gemeinschaft als angesehene Position gilt.

3.6.3 Algorithmische Sichtbarkeit

Top-Kommentare werden nicht zufällig angezeigt, sondern durch die Like-Funktion und weitere unbekannte Faktoren vom YouTube-Algorithmus bestimmt. So spiegeln sie die am stärksten bestätigten Meinungen, nicht die am häufigsten vorkommenden wider (Möller et al., 2021). Das verstärkt möglicherweise eine Form der Schweigespirale, bei der sich Personen eher äußern, die sich von der öffentlich wahrgenommenen Meinung bestätigt fühlen (Noelle-Neumann, 1974). Ein weiterer wichtiger Faktor ist, dass Videoersteller die Kommentare unter ihren Videos moderieren können. Es ist möglich, bestimmte Begriffe automatisch zu filtern, Nutzer auf einem Kanal auszublenden oder Kommentare direkt zu löschen. So können unerwünschte Kommentare entfernt und der Diskurs beeinflusst werden. Auch kann so jede Erwähnung von KI systematisch verstummt werden lassen, obwohl eine mögliche Mehrheit der Kommentierenden genau dies anspricht. Um solche Fälle bestmöglich zu umgehen, wurden Kanäle und deren Videos, die keinerlei Kommentare zum Thema hatten, bei der Suche ausgeschlossen. Um zu testen, ob ein Kanal solche Filter aktiv hat, kann ein Kommentar mit dem vermuteten Begriff geschrieben werden und auf einem anderen Account nach diesem Kommentar geschaut werden, wenn er nicht zu sehen ist, ist es sehr wahrscheinlich, dass Filter aktiv sind. Dennoch führen die zuvor genannten Punkte dazu, dass der Kommentarkorpus kein völlig neutrales Abbild der Rezeption ist, sondern nur ein algorithmisch und sozial geformter Ausschnitt.

3.6.4 Subjektivität in der Kodierung

Die Kodierung wurde von einer einzelnen Person durchgeführt. Im Rahmen der Arbeit war es nicht möglich, eine Interrater-Reliabilitätsprüfung durchzuführen, bei der mehrere Personen das Material kodieren, um zu prüfen, ob das Kategoriensystem zuverlässig ist und nicht nur von einer Person nachvollziehbar. Besonders bei Grenzfällen kommt die Subjektivität des Kodierenden zum Tragen, wenn ein Kommentar zwischen zwei Kategorien liegt oder die Valenz ambivalent ist. Als Gegenmaßnahme wurde ein Kodierbuch mit expliziten Definitionen, Ankerbeispielen und Abgrenzungsregeln erstellt, um die Nachvollziehbarkeit und Konsistenz über den Kodierprozess hinweg zu sichern. Um temporäre Urteilsverzerrung zu verhindern, wurde in mehreren Durchläufen kodiert (Erstkodierung, Revision, Feinkodierung). Dennoch bleibt die Kodierung interpretativ und die Ergebnisse als qualitative Befunde, nicht als objektive Messung zu betrachten.

3.6.5 Grenzen der Zuordnung von Stimmfunktionen

Die Kategorien sind theoriegeleitet aus der Literatur hergeleitet, die nicht für YouTube-Kontexte entwickelt wurde, die Übertragung ist also eine konzeptionelle Eigenleistung mit Interpretationsspielraum. Die Funktion einer Stimme kann sich innerhalb eines Videos temporär verändern. So kann ein Video, das als Hauptfunktion die Stimme als Werkzeug einordnet, zwischenzeitlich die Stimme als Erzählstimme nutzen. Die Zuordnung beeinflusst so auch, welche Kommentare der jeweiligen Stimmfunktion zugerechnet werden. Fehlzuordnungen könnten direkte Auswirkungen auf die Vergleichbarkeit der Befunde haben.

3.6.6 Alternative Erklärungen für Muster im Material

Für Befunde, die im Ergebnisteil gefunden werden, gibt es oft auch andere mögliche Erklärungen, da bei YouTube-Kommentaren als Daten mehr Variablen relevant sind als festgestellt werden können. Daher kann ein Ergebnis nicht explizit und ausschließlich einem bestimmten Umstand zugeordnet werden. Unterschiede zwischen Funktionskategorien könnten durch Kanalcharakteristika erklärt werden statt durch Stimmfunktion. Zum Beispiel könnten Persona-Kanäle tendenziell kleiner mit engagierteren Communities sein. Auch Thema und Inhalt der Videos können Unterschiede erklären, so kann KI-Nutzung bei KI-nahen Inhalten positiver betrachtet werden als bei künstlerischen Themen. Unterschiede könnten auch durch Kanalkenntnis der Kommentierenden beeinflusst werden. Stammzuschauer eines Kanals könnten anders auf KI-Stimmen reagieren als neue Zuschauer. Diese Alternativerklärungen können im vorliegenden Design nicht ausgeschlossen werden und müssen bei der Interpretation der Ergebnisse mitgedacht werden.

4 Ergebnisse

4.1 Beschreibung des Korpus

Der Korpus umfasst insgesamt 87 Videos von 30 Kanälen, jeweils zehn pro Stimmfunktion. Erhoben wurden 862 Top-Kommentare sowie 1.044 als relevant eingestufte Kommentare, die sich explizit auf die Stimme oder den KI-Einsatz und nicht klar auf KI-Einsatz in anderer Form, zum Beispiel das Skript oder visuelle Inhalte, beziehen. Auffällig ist, dass die Werkzeug-Funktion trotz der geringsten Aufrufzahl mit 426 relevanten Kommentaren deutlich mehr stimmbezogene Reaktionen auslöste als Erzählstimme (323) und Persona (295) (siehe Tabelle 1). Und das, obwohl die Videos der Stimme als Werkzeug insgesamt mit 32,9 Millionen Aufrufen am wenigsten hatten im Vergleich zu den 37,2 Millionen Aufrufen der Videos von Erzählstimme und 34,9 Millionen Aufrufen von Persona. Die geringfügigen Abweichungen von der angestrebten Videoanzahl sind auf einzelne Kanäle zurückzuführen, die zum Erhebungszeitpunkt weniger als drei geeignete Videos aufwiesen. Aufgrund von zum Erhebungszeitpunkt nicht verfügbaren oder bereits gelöschten Kommentaren weicht die tatsächliche Anzahl der Top-Kommentare in einzelnen Fällen von der angestrebten Höchstzahl von zehn pro Video ab.

Tabelle 1
Übersicht des Analysekorpus nach Stimmfunktion

Stimmfunktion	Kanäle	Videos	Top-Komm.	Relevante Komm.	Aufrufe	KI-Kennz. (Kanäle)	KI-Kennz. (Videos)
Werkzeug	10	30	297	426	32,9 Mio.	1/ 10	3/ 30
Erzählstimme	10	29	285	323	37,2 Mio.	2/ 10	6/ 29
Persona	10	28	280	295	34,9 Mio.	4/ 10	12/ 28
Gesamt	30	87	862	1044	105 Mio.	7/ 30	21/ 87

Anmerkung. Aufrufe gerundet auf 0,1 Mio. KI-Kennzeichnung bezieht sich auf die explizite Angabe des KI-Einsatzes durch den Kanal. Die geringfügigen Abweichungen von der angestrebten Videoanzahl sind auf einzelne Kanäle zurückzuführen, die zum Erhebungszeitpunkt weniger als drei geeignete Videos aufwiesen.

Die relevanten Kommentare verteilen sich auch unterschiedlich innerhalb der Funktionskategorien auf Kanäle, die KI kennzeichnen und solche, die es nicht tun. Die Kennzeichnung des KI-Einsatzes durch den Kanal verteilt sich über die drei Funktionskategorien sehr ungleich. Während bei Persona fast die Hälfte der Kanäle KI explizit angeben, ist es bei Werkzeug nur einer von zehn (siehe Tabelle 1). Gleichzeitig enthalten bei der Stimme als Werkzeug 361 von 426 relevanten Kommentaren (85 %) einen expliziten KI-Bezug, bei der Stimme als Persona 211 von 295 (72 %) und bei der Stimme als Erzählstimme 103 von 323 (32 %). Werkzeug-Videos kennzeichnen bei den für die Untersuchung ausgewählten Videos KI am seltensten, lösen aber gleichzeitig am häufigsten explizit KI-bezogene Kommentare aus. Es ist zu

beachten, dass im Rahmen dieser Untersuchung keine Aussagen aus den Zahlen geschlossen werden sollten. Diese werden lediglich zur Beschreibung des Korpus herangezogen und erlauben keine Rückschlüsse auf die Repräsentativität der Reaktionen oder als Grundlage für inhaltliche Schlussfolgerungen.

Für die Analyse der Daten sind vier Punkte zu beachten. Einzelne Kanäle sind überrepräsentiert, da sie mehr relevante Kommentare enthalten, während andere unterrepräsentiert sind. Es können also auch aus diesem Grund keine Verallgemeinerungen gemacht werden. Außerdem enthält die Stimme in Funktion der Persona zwei verschiedene Kanaltypen (V-Tuber und Avatar-Kanäle), die sich bezüglich der Reaktionen in den Kommentaren stark voneinander unterscheiden. Bei einzelnen Kanälen, die Stimme als Erzählstimme nutzen, wurde vermehrt die thematische Inkongruenz zwischen Videoinhalt und KI-Einsatz angemerkt, in diesen Fällen sind negative Reaktionen besser mit dem Inhalt als mit der Stimmfunktion zu erklären. Bei einigen Kanälen ist KI angegeben und bei anderen nicht, was in der Auswertung beachtet werden muss und einen direkten Vergleich erschwert.

4.2 Rekonstruktion der Auswertungskategorien

4.2.1 Tragfähigkeit der Kategorien

Alle vier Hauptkategorien haben sich im Material belegen lassen. Keine der Kategorien musste verworfen werden. Natürlichkeit und Klang (NK) ist tragfähig und differenziert sich klar in vier Unterkategorien. Die Klangqualität wird von Zuschauern auf folgenden Ebenen kommentiert: Ausdruck, Sprachfehler, positiver Eindruck und Wiedererkennung. Glaubwürdigkeit und Vertrauen (GV) sind im Material seltener thematisiert. Authentizität und Identität (AI) ist ebenfalls tragfähig und wird sowohl positiv als auch negativ wahrgenommen. Reaktion auf KI (RK) bildet im Material die größte Kategorie und an vielen Stellen auch die Basis für andere Kategorien.

4.2.2 Unterkategorien

Zu den vier deduktiv gebildeten Hauptkategorien wurden aus dem Material heraus 20 Unterkategorien und eine Restkategorie für nicht relevante Kommentare (NR) gebildet.

Für Natürlichkeit und Klang ließen sich aus dem Material vier Unterkategorien definieren: Fehlender Ausdruck und Monotonie (NK.1), Technische Artefakte und Aussprachefehler (NK.2), Positiver Klangeindruck (NK.3), Wiedererkennung der Stimme (NK.4). NK.1 erfasst Kommentare, die die Stimme als emotionslos oder ausdrucksarm beschreiben, ohne konkrete Fehler zu benennen, während NK.2 sich auf konkrete Fehler in der Sprache, wie Aussprachefehler, falsches Tempo oder andere technische Artefakte bezieht. NK.3 entstand, um explizit positive Bewertungen des Klangs in Abtrennung zur Abwesenheit von negativer Kritik zu trennen. Da auf voneinander unabhängigen Kanälen durch KI häufig die gleiche Stimme verwendet wird,

kommt es in den Kommentaren zur Wiedererkennung eben dieser Stimmen, um diese Kommentare zu erfassen, wurde NK.4 definiert.

Für Glaubwürdigkeit und Vertrauen ließen sich drei Unterkategorien definieren: Zweifel an der inhaltlichen Richtigkeit (GV.1), Erkennen des Scripts als KI-generiert (GV.2) sowie Vertrauen in den Inhalt trotz erkanntem KI-Einsatz (GV.3). GV.1 erfasst Kommentare, die die sachliche Richtigkeit der Videos aufgrund der Nutzung von KI anzweifeln. Durch GV.2 werden hingegen Kommentare erfasst, die beispielsweise aufgrund von Redundanz das Skript, welches ein wichtiger Teil der Stimmerscheinung ist, als KI-typisch identifizieren und bemängeln. Um Kommentare zu erfassen, die konkret KI-Inhalte des Videos kritisieren, aber den Videoinhalt trotzdem loben, wurde GV.3 definiert.

Für Authentizität und Identität ließen sich fünf Unterkategorien definieren: die Forderung nach einer echten menschlichen Stimme (AI.1), die Akzeptanz der synthetischen Stimme als Teil einer Figur oder Kanalidentität (AI.2), die irrtümliche Zuschreibung der Stimme als menschlich (AI.3), die Thematisierung der Stimme als Wiedererkennungsmerkmal des Kanals (AI.4) sowie parasoziale Enttäuschung nach dem Erkennen des KI-Einsatzes (AI.5). Unabhängig von Klang erfasst AI.1 die Forderung nach einer menschlichen Stimme. Speziell zur Erfassung der Akzeptanz einer Stimme als zugehörig zur Kanalidentität wurde AI.2 gebildet. Die synthetische Stimme wird in dieser Unterkategorie nicht als Defizit, sondern grundlegendes Merkmal einer Figur und Kanalidentität wahrgenommen. AI.3 erfasst positive Bewertungen der Stimme, die darauf basiert, dass die Stimme irrtümlich von Kommentierenden als menschlich eingestuft wurde. Diese Kategorie unterscheidet sich von NK, da die Bewertung hierbei auf der Fehlzuschreibung beruht. AI.4 entstand, weil Kommentare die Kontinuität der Kanalstimme thematisierten. Bei diesen Kommentaren handelt es sich nicht um Forderung nach menschlicher Stimme, sondern um das Hinterfragen der Kanalidentität. AI.5 wurde definiert, da sich emotionale Enttäuschung nach KI-Erkennung von reiner Ablehnung unterscheidet. Hierbei waren Kommentierende zuvor davon ausgegangen, die Inhalte wären nicht mit KI erstellt und beschrieben nach Erkennung der Stimme als KI konkrete Abweisung oder Enttäuschung.

Für die Reaktion auf KI-Einsatz ließen sich neun Unterkategorien definieren: Unsicherheit über den KI-Status (RK.0), pauschale Kurzablehnung (RK.1), Plattformkritik (RK.2), Transparenzforderung an den Creator (RK.3), ambivalente Akzeptanz (RK.4), inhaltliche Inkongruenz und Ironie (RK.5), Tool-Interesse und Werkzeuganfragen (RK.6), Abwendungsverhalten (RK.7) sowie positive Akzeptanz und Befürwortung (RK.8). RK.0 wurde nachträglich induktiv gebildet, weil ein großer Teil der Kommentare weder Ablehnung noch Akzeptanz konkret ausdrückte, sondern Unsicherheit über den KI-Status des Videos. RK.1 erfasst pauschale Ablehnung von KI ohne eine Begründung. Diese Form der Reaktion hat eine eigene Kategorie, da sie zu unterscheiden ist von begründeter Kritik, um differenzierter nachvollziehen zu können, welche möglichen Gründe Kritik hat und dementsprechend, wie Videoersteller damit umgehen

können. RK.2 umfasst Plattformkritik, wie fehlende KI-Filtermöglichkeiten oder Kennzeichnung. In Abgrenzung daran erfasst RK.3 die Forderung der Kennzeichnung durch den Videoersteller. Diese beiden Kategorien sind getrennt, um klar zu erfassen von wem Kommentierende Änderung erwarten. Da sich in den Kommentaren zeigte, dass die gleichzeitige Ablehnung von KI-Nutzung und Akzeptanz im Einzelfall eine eigenständige Haltung bildet, entstand RK.4. Im Unterschied zu GV.3 bezeichnet RK.4 die konkrete Akzeptanz im Einzelfall, nicht das Vertrauen in KI-generierte Inhalte. Um Kommentare zu erfassen, die den Widerspruch zwischen Videoinhalten und KI-Einsatz als Ironie thematisierten, jedoch keine Klangkritik oder pauschale Ablehnung beschreiben, wurde RK.5 gebildet. RK.6 erfasst das Interesse an KI-Nutzung in Form von Werkzeugfragen ohne Wertung, da diese sich als eigenständige Reaktionsform herausstellten. In Abgrenzung zu RK.1 beschreiben Kommentare, die RK.7 zugeordnet wurden, konkretes Ablehnungsverhalten, in Form von Ankündigungen das Video abzuschalten, zu disliken und ähnlichem. Um das gesamte Reaktionsspektrum abzubilden, wurde RK.8 definiert, wodurch Kommentare erfasst werden, die KI-Nutzung positiv bewerten.

Tabelle 2 gibt eine kompakte Übersicht aller Haupt- und Unterkategorien mit Kurzdefinition und Ankerbeispiel. Das vollständige Kodierbuch mit Abgrenzungsregeln findet sich in Anhang A.

Tabelle 2
Übersicht des Kategoriensystems (Unterkategorien)

Kategorie	Code	Kurzdefinition	Ankerbeispiel
Fehlender Ausdruck	NK.1	Stimme als emotionslos / monoton beschrieben	„dead bland computer generated voice“
Technische Artefakte	NK.2	Konkrete Fehler: Aussprache, Tempo, Pausen	„AI voice can't tell the difference between resume and résumé“
Positiver Klangeindruck	NK.3	Stimme explizit als angenehm bewertet	"your voice is so calming and gentle"
Stimme wiedererkannt	NK.4	Bekannte TTS-Stimme von anderen Kanälen erkannt	"The same AI Generated voice yet again!!!"
Inhalt bezweifelt	GV.1	Faktentreue wegen KI angezweifelt	"Has this been fact checked? Or is it just AI distortion?"
Script als KI erkannt	GV.2	Stil des Scripts als KI-typisch identifiziert	"This script is so obviously AI generated"
Vertrauen trotz KI	GV.3	Inhalt akzeptiert trotz erkanntem KI-Einsatz	"SO this is AI, right? Good analysis though"

Menschl. Stimme gefordert	AI.1	Echte menschliche Stimme als Norm gefordert	"Does anyone narrate anymore?"
Persona-Akzeptanz	AI.2	TTS als Teil der Kanalidentität akzeptiert	"You have a lovely robot voice"
Irrtüml. Menschzuschreibung	AI.3	Stimme fälschlich als menschlich eingestuft	"THANK YOU for no AI!!!"
Kanalidentität hinterfragt	AI.4	Unsicherheit, ob Stimme noch zum Creator gehört	"your voice is completely changed. Are you old byl?"
Parasoziale Enttäuschung	AI.5	Emotionaler Verlust bei KI-Erkenntnis	"it sucks all the joy I had for 15 minutes out"
Unsicherheit	RK.0	Unsicherheit, ob Stimme/Video KI-generiert ist	"Is this AI?"
Kurzablehnung	RK.1	Pauschale Ablehnung ohne Begründung	"AI slop"
Plattformkritik	RK.2	Forderung an YouTube nach Filterung	"AI videos should be flagged"
Transparenzforderung	RK.3	Forderung an Creator nach Kennzeichnung	"u should mark this as AI"
Ambivalente Akzeptanz	RK.4	Grundsätzl. Ablehnung, aber Einzelfall ok	"I REALLY dislike AI but Loved this!"
Inhaltl. Inkongruenz	RK.5	Widerspruch zwischen Inhalt und KI-Einsatz	"The irony of AI narration in a piece on 'made to last'"
Tool-Interesse	RK.6	Frage nach dem verwendeten TTS-Tool	"What AI text to speech do you use?"
Abwendungsverhalten	RK.7	Ankündigung wegzuklicken/zu disliken	"The second I hear the AI voice I click off"
Positive Akzeptanz	RK.8	KI-Einsatz explizit begrüßt	"This is what AI was made for"

Anmerkung. Hauptkategorien deduktiv gebildet und durch Farben markiert. Natürlichkeit und Klang, Glaubwürdigkeit und Vertrauen, Authentizität und Identität und Reaktion auf KI. Unterkategorien induktiv aus dem Material gebildet. Mehrfachkodierung möglich. Das vollständige Kodierbuch mit Abgrenzungsregeln findet sich in Anhang A.

4.2.3 Bedeutung und Gewichtung der Kategorien im Material

RK war im Material am stärksten vertreten, konkret weil viele Kommentare unter RK.0 und RK.1 fielen, also Unsicherheit über KI-Nutzung ausdrückten oder pauschal KI ablehnten. RK.1 ist die häufigste Einzelreaktion. Über alle Funktionskontexte hinweg erwies sich die Unsicherheit über KI-Nutzung (RK.0) als zentral. Ähnlich stark wurde die Forderung nach menschlicher Stimme (AI.1), auch funktionsübergreifend, ausgedrückt. Speziell in der Persona-Kategorie wurde die Akzeptanz der Stimme als Teil der Kanalidentität (AI.2) verstärkt ausgedrückt. Bezogen auf den Klang erwies sich die NK.1 als die am häufigsten auftretende und somit sehr relevante Unterkategorie. Randständig waren die Kategorien RK.2, RK.3, RK.4 und RK.8. Auch AI.3 und GV.3 treten nur vereinzelt auf.

Für den Vergleich ist RK.1 als Ausgangspunkt und dominante Grundreaktion wichtig, um zu verstehen, ob die pauschale Ablehnung sich zwischen den Stimmfunktionen verschieden stark äußert. Auch NK als Hauptkategorie ist für den Vergleich relevant, da Klangkritik in den drei Funktionen unterschiedlich stark auftritt, selbst wenn die technische Qualität gleichwertig ist. Auch AI.1 und AI.2 sind durch ihre entgegengesetzte Ausrichtung innerhalb der gleichen Hauptkategorie wichtig, um die Authentizitätswahrnehmung vergleichen zu können. RK.5 tritt fast ausschließlich in einem Funktionskontext auf und zeigt, dass einige Reaktionen möglicherweise stärker durch das Videothema als die Stimmfunktion beeinflusst werden. Anders als RK.5 tritt RK.6 in verschiedenen Funktionskontexten und aus verschiedenen Gründen auf, deshalb ist es wichtig, hier innerhalb der Unterkategorie inhaltlich stärker zu vergleichen. Auch die Heterogenität der Persona-Funktion, die durch den Unterschied zwischen VTuber- und KI-Avatar-Kanal entsteht, ist bei Vergleichen zu beachten.

4.3 Stimme als Werkzeug

4.3.1 Kommentierung der Stimme als Werkzeug

Bei Videos, in denen die Stimme die Werkzeug-Funktion einnimmt, dominiert die pauschale Kurzablehnung (RK.1). Kommentare wie „AI slop“ (K004_V1_S02, K025_V1_S17, K007_V2_S04, K024_V3_S18), „Fucking Ai bot voiceover“ (K030_V1_S03) oder “AI voice garbage.” (K024_V3_S02) zeigen, dass die Erkennung von KI-Nutzung ausreichender Grund zur Ablehnung, ohne inhaltliche Auseinandersetzung, ist. Viele Kommentare fragen allerdings auch explizit nach der KI-Stimme, die genutzt wird (K017_V2_S18, K017_V2_S16, K006_V1_S06), gerade bei Tutorial-Kanälen, in deren Publikum selbst viele Videoersteller sind. In einigen Kommentaren, die RK.7 zugeordnet sind wird auch ohne emotionale Begründung sachlich das Abwendungsverhalten beschrieben. Solche Kommentare sind zum Beispiel „The second i hear the ai voice i click off“ (K007_V3_S01), „not even going to watch this, AI voice the entire time“ (K025_V1_S03) und „Disliked because fully ai“ (K025_V1_S11). In dieser

Funktion wird die Stimme auch vermehrt mit Inhaltszweifeln (GV.1) verknüpft. Kommentare sagen Videos enthalten Falschinformationen, die durch simple Faktenchecks verhindert werden könnten (K025_V1_S08, K023_V1_S04). In Einzelfällen wird die Stimme auch wiedererkannt (NK.4). Ein Kommentar beschreibt „Im so sick of this voice“, er zeigt also Ablehnung gegenüber der Stimme, weil er sie zu oft gehört hat. Kommentare fordern bei den meisten der Videos mit dieser Stimmfunktion eine menschliche Stimme (K002_V1_S05, K004_V2_S05, K025_V2_S11). In vielen Fällen fallen diese Kommentare auch positiv ermutigend aus (K004_V3_S03). Besonders bei einem Kanal fällt vielen Kommentierenden die falsche Aussprache der KI-Stimme auf (K025_V3_S04, K025_V3_S05, K025_V3_S09). Bis auf Ausnahmen befassen sich die gesammelten Top-Kommentare bei dieser Funktion nicht mit der Nutzung einer KI-Stimme.

4.3.2 Kriterien positiver, negativer und neutraler Rahmung

In diesem Kontext wurde die Stimme vor allem dann positiv bewertet, wenn sie unauffällig und entspannend war. Kommentare beschreiben hier „your voice is so calming“ (K024_V1_S25) und „i love your voice, it’s so calming and gentle“ (K024_V3_S03), wobei beide diese Kommentare auch darauf hinweisen, dass eine irrtümliche Menschzuweisung (AI.3) stattfindet und dies ein Faktor für die positive Bewertung sein könnte. Auch finden sich viele Kommentare, die Ambivalenz (GV.3) ausdrücken und die Inhalte als positiv, die Nutzung von KI aber als negativ bewerten (K025_V1_S23, K030_V1_S10). Wobei ein Kommentar allerdings auch darauf hinweist, dass inhaltlich wahrgenommene Qualität Stimm-Ablehnung mildert (K025_V3_S22). Neutral wird die Stimme fast ausschließlich im Rahmen des Interesses an genutzten Werkzeugen bewertet (K006_V2_S01), während die Nutzung von KI allein in vielen Fällen schon als Kriterium negativer Rahmung ausreicht (K004_V1_S02).

4.3.3 Rolle von Verständlichkeit, Effizienz und Reibung

Besonders bei dem Kanal BoNa (K025) zeigte sich, dass Verständlichkeit in dieser Funktion eine wichtige Rolle spielt. Viele Kommentare beschweren sich über fehlerhafte Aussprache (K025_V3_S04), was in manchen Fällen dazu führt, dass dies Grund zu Ablehnungsverhalten (RK.7) wie dieser Kommentar beschreibt: „The AI. Mispronouncing words. I. Can't. Continúe“ (K025_V3_S27). Auch zu hohes Tempo wird kritisiert, welches dazu führt, dass Kommentierende dem Gesagten nicht folgen können (K029_V1_S03). In diesem Fall wird die KI-Stimme selbst akzeptiert und lediglich Geschwindigkeit und somit Verständlichkeit kritisiert. Indikator für ausbleibende Reibung sind auch die häufig gestellten Fragen bezüglich der genutzten Stimme (K023_V2_S20), da diese direkt oder indirekt ausdrücken, dass die genutzte Stimme gut verständlich ist und keine Reibung verursacht. Es gibt auch einzelne Hinweise, dass sich Klangreibung und Inhaltskritik gegenseitig verstärken (K023_V1_S04). Insgesamt wird deutlich, dass in der Werkzeug-Funktion nicht nur die KI-Nutzung als solche, sondern die

technische Ausführung, besonders Aussprache und Tempo, ausschlaggebend für die Akzeptanz der Stimme ist.

4.3.4 Relevanz von Authentizität und Funktionalität

Vereinzelt fordern Kommentare eine menschliche Stimme (K002_V1_S05) und wundern sich, ohne dies zu bewerten, über die Änderung der Stimme innerhalb eines Kanals (K006_V1_S08). Das deutet darauf hin, dass einige Zuschauer bei Stimmen in dieser Funktion Authentizitäts- und Kontinuitätserwartung haben, aber keine konkrete parasoziale Bindung aufweisen. Ein Kommentar beschreibt „Script By Chatgpt 😇 Voice by ElevenLabs 😇 Always listen to those with real experience, not to someone generating script or using an AI voice“ (K030_V2_S03), was zeigt, dass durch eine menschliche Stimme gesteigerte Authentizität zu generell höherer Akzeptanz und positiverer Bewertung der Inhalte im Ganzen führen kann. Insgesamt zeigt sich, dass die Funktionalität der Stimme stärker im Fokus steht als die Stimme selbst. Kritik richtete sich häufiger gegen die KI-Nutzung insgesamt als gegen die Stimme.

4.3.5 Relevanz von KI-Kennzeichnung

Im gesammelten Material der Stimmfunktion Werkzeug spielt die KI-Kennzeichnung keine große Rolle, da nur ein Kanal diese Angabe macht und der größte Teil der Kommentare aus nicht gekennzeichneten Videos stammt. Die Kennzeichnung beim Kanal Zeven 7 (K029) erzeugte hauptsächlich mehr Interesse an genutzten Programmen (RK.6). Die gesammelten Kommentare von diesem Kanal beschrieben keine Unsicherheit über KI-Inhalte, während das bei Videos ohne Kennzeichnung vorkam.

4.4 Stimme als Erzählstimme

4.4.1 Anforderungen an die Stimme

Bei Videos mit Erzählstimme ist die Forderung nach einer menschlichen Stimme sehr stark vertreten. Kommentare wie „Rest in peace, the days when actual humans used their actual human voices to convey real emotion“ (K008_V1_S22) drücken Frustration über KI-Stimmen im Allgemeinen aus, da diese keine echten Emotionen ausdrücken können. Es zeigt sich auch, dass Kommentierende die Stimme als zu monoton ohne Ausdruck empfinden (K008_V1_S39). Die Stimme soll nicht nur verständlich, sondern auch emotional und ausdrucksstark präsentieren. Die pauschale Ablehnung von KI (RK.1) findet auch hier statt. Besonders bei eher wissenschaftlichen Themen, aber auch verteilt über Videos dieser Funktion findet die Forderung nach transparenter Kennzeichnung der KI-Inhalte (RK.3) statt (K003_V1_S46). Diese Forderung entsteht unter anderem aus der Unsicherheit, ob es sich um eine KI-Stimme handelt, wie dieser Kommentar beschreibt: „I can't tell if the narrator is Ai and that's bothering me. If it's not can you add 'Not an Ai' on the description“ (K003_V1_S46). Auch die Unsicherheit über KI-

Nutzung (RK.0) als solches ist sehr präsent und zeugt indirekt auch von gewünschter Transparenz (K009_V1_S16). Positiv wird hier besonders häufig der Klang der Stimme bewertet (K010_V1_S01). Dies tritt vor allem beim Kanal Gates of Imagination (K010) auf, welcher den genutzten KI-Stimmen Namen zuweist. Das könnte eine irrtümliche Menschzuweisung (AI.3) verursachen, was sich in den Kommentaren die, die Stimme positiv bewerten zum Beispiel durch die direkte Anrede „your voice“ (K010_V1_S07) zeigt. Die gesammelten Top-Kommentare kommentieren die Stimme in dieser Funktion, bis auf einer Ausnahme, nicht.

4.4.2 Authentizität im narrativen Kontext

Bei der Erzählstimme tritt häufig AI.5 auf. Ein Kommentar beschreibt: „The voice is AI … it sucks all the joy I had for 15 minutes out“ (K003_V1_S23). Das zeigt eine parasoziale Enttäuschung, die auftritt, wenn Zuschauer erst nach einer gewissen Dauer des Konsums herausfinden, dass die Stimme nicht menschlich ist. Diese Enttäuschung und das Gefühl betrogen worden zu sein beschreibt auch der Kommentar „This whole video (and channel) AI generated?? Wow I got suckered in!“ (K009_V1_S10) sehr deutlich. Dabei geht es nicht um bloße Klangkritik, sondern die Wahrnehmung vorgetäuschter Authentizität, die durch den Nutzen von KI nicht mehr besteht. Auch häufige Fehlzuweisungen (AI.3) bei dieser Stimmfunktion, die explizit sagen „Enhanced by not using an Ai voice“ (K012_V1_S12) zeigen, dass Kommentierende eine Menschenstimme erwarten und Authentizität in Form einer menschlichen Stimme gefordert wird. Auch auf Skript-Ebene wird Authentizität gefordert. Kommentierende zweifeln die Authentizität der Skripte an, die sie als KI-generiert identifizierten (K022_V1_S02). In dieser Funktion ist die Frage, ob ein Mensch hinter den Inhalten steckt, besonders präsent.

4.4.3 Rolle von Passung und Glaubwürdigkeit

Ein Extremfall positiver Passung bildet Gates of Imagination (K010). Fast alle positiven Klangbewertungen (NK.3) im Material dieser Stimmfunktion kommen von diesem Kanal. Das Hörbuch-Format erzeugt Erwartungen, die von der Stimme erfüllt werden, wobei der KI-Einsatz kaum thematisiert wird. Hier erkennen Kommentierende die Stimmen aber auch seltener als solche, was zu einer Verzerrung der Bewertung führen kann. Außerdem legt das Format Hörbuch einen größeren Fokus auf die Stimme, sodass sie häufiger kommentiert wird. Einzelne Kommentare beschreiben aber auch die Akzeptanz und gute Passung der Stimme trotz erkannter KI-Nutzung (K010_V3_S13), demnach können gut klingende Stimmen im Hörbuch-Kontext ambivalente Akzeptanz (RK.4) auslösen. Der gegenteilige Effekt kann jedoch auch erzeugt werden, wenn die Stimme vom Inhalt ablenkt (K009_V3_S06). Sie kann als Störfaktor wirken, der Zuwendung zu Inhalten verhindert (K009_V1_S20). Vereinzelte Kommentare beschreiben gleichermaßen Inhaltszweifel aufgrund der KI-Nutzung (GV.1) als auch Vertrauen in die Glaubwürdigkeit trotz erkannter KI-Nutzung. Unglaubwürdigkeit entsteht, wenn die KI-Stimme und Videoinhalt nicht zusammenpassen (RK.5). Es werden Zweifel an der

Glaubwürdigkeit und dem Wissensstand der KI-Stimme geäußert, da diese keine eigenen Erfahrungen haben kann, die ihr die Autorität geben würden, Aussagen zu bspw. physischer Arbeit zu machen (K013_V2_S12). Diese Art von Ausdruck der Unglaubwürdigkeit findet sich innerhalb der Funktion nur in Videos, die sich thematisch mit der realen Welt befassen und nicht nur Geschichten erzählen. Konkret der Kanal Industrial Decay (K013) wird diesbezüglich stark kritisiert. Dabei wird besonders die Ironie der KI-Nutzung beim Thema der Qualität alter Werkzeuge thematisiert (K013_V1_S11).

4.4.4 Wahrnehmung von Künstlichkeit der Stimme

RK.5 ist fast ausschließlich im Material zur Erzählstimme vertreten und entsteht durch thematische Inkongruenz. Die Künstlichkeit stört hier nicht im Klang der Stimme, sondern wird als autoritätsmindernder Faktor wahrgenommen. Dieses Muster tritt vornehmlich bei Videoinhalten auf, die eine negative Grundhaltung gegen neue Technik, wie bei Industrial Decay (K013), oder KI haben, wie bei MonkeyExplains (K019). Auch durch konkrete technische Fehler (NK.2) kann eine durch die Künstlichkeit verursachte Störung bei Zuschauern verursacht werden. Kommentare beschreiben, dass die KI-Stimme nicht atmet. Natürliche Atempausen und natürlicher Rhythmus werden bei einer Erzählstimme als Grundvoraussetzung wahrgenommen. Auch äußert ein Kommentar, dass die KI-Stimme als Symbol für inhaltliche Oberflächlichkeit wahrgenommen wird (K013_V3_S07). Eine distanzierte Erzählhaltung wie bei Gates of Imagination verlangt in manchen Fällen im Gegensatz dazu sogar eine monotone Stimme, wie ein anderer Kommentar beschreibt (K010_V3_S09).

4.4.5 Relevanz von KI-Kennzeichnung

Die Kennzeichnung von KI führt hier dazu, dass die Unsicherheit (RK.0) in den Kommentaren geringfügig reduziert, jedoch nicht vollständig überwunden wird. Die parasoziale Enttäuschung (AI.5) tritt jedoch bei Kanälen mit KI-Kennzeichnung im Material deutlich häufiger auf. Bei Kanälen mit Kennzeichnung traten in der Untersuchung zudem seltener Transparenzforderungen (RK.3) auf. Doch auch bei den gekennzeichneten Videos gab es einen Kommentar, der explizitere Kennzeichnung forderte: „Listen it is educational, but u should mark this as AI. People barely read the description box“ (K019_V1_S09). Im Material tritt positive Akzeptanz (RK.8) ausschließlich bei gekennzeichneten Videos vom Kanal Cat Lovers Forum (K008) auf.

4.5 Stimme als Persona

4.5.1 Erwartungen an Persönlichkeit, Echtheit und Sprecheridentität

Die Kategorie wird strukturell in zwei Untergruppen mit fast gegensätzlichen Erwartungsmustern geteilt. Die erste Gruppe besteht aus VTubern (Zentreya, Neuro-sama, MOTHERv3) und die zweite aus Avatar-Kanälen (AI Guys, BigStepsMedia, Bioforceman, Janxt, Yellow Dude,

Captain Workout, Ironvayne). Die beiden Gruppen unterscheiden sich in der Wahrnehmung. Dabei fällt auf, dass Avatar-Kanäle anders als VTuber die sekundäre Funktion Werkzeug haben, was sich darin widerspiegelt, dass ihre Kommentare stärker der Werkzeug-Funktion als der Persona-Funktion ähneln.

4.5.1.1 Avatar-Kanäle

Bei Avatar-Kanälen dominiert die pauschale Ablehnung von KI-Nutzung (RK.1) und es ist keine funktionsspezifische Persona-Erwartung erkennbar. Bezogen auf den Klang ist fehlender Ausdruck bei diesen Kanälen die häufigste Kritik. Einige beschreiben sogar, dass versucht wird, menschlich zu klingen, um besser anzukommen (K028_V3_S11). Ein Stimmwechsel wirkt auf Kommentierende gerade dann irritierend, wenn sie sich nicht sicher sind, ob es sich um eine KI-Stimme handelt (K028_V2_S03). Auch wird der Wechsel von einer menschlichen zu einer KI-Stimme kritisiert und sogar hinterfragt, ob noch die gleiche Person hinter dem Kanal steht (K028_V3_S17).

Bei Avatar-Kanälen wird die Stimme als Ersatz menschlicher Präsenz wahrgenommen und dafür kritisiert (K001_V2_S03, K014_V1_S10). Besonders Videos mit Fitnessinhalten werden für fehlende Identität und Echtheit kritisiert. Dabei sehen Kommentare die Stimme und den Avatar nicht als Sprecheridentität oder Person hinter dem Video, sondern zweifeln die Integrität des gesamten Kanals an (K005_V1_S03). Auch hier wird Transparenz über die KI-Nutzung (RK.3) gefordert und Frustration über ein vorgetäuschtes Ich geäußert (K014_V2_S01).

4.5.1.2 VTuber-Kanäle

Bei VTubern hingegen wird die Stimme als Teil der Persönlichkeit akzeptiert (AI.2) und als Teil der Figur diskutiert (K021_V1_S02). Die robotisch klingende Stimme wird sogar als positiver Teil, der die Figur ausmacht, bezeichnet (K021_V2_S02). Bei VTubern wird fehlender Ausdruck als charakteristisches Merkmal der Figur akzeptiert. Im Material wird die Nutzung von KI-Stimmen in dieser Untergruppe auch als Maskierung der echten Stimme akzeptiert (K027_V2_S15) und die Kommentare äußern eher Neugier als Ablehnung (K027_V2_S17).

4.5.2 Authentizität in der Persona-Funktion

Die Kategorie Authentizität und Identität (AI) wird im Material dieser Funktion am stärksten diskutiert. Authentizitätskritik und -akzeptanz sind nahezu gleichermaßen vertreten. Die Akzeptanz war bei der Untergruppe VTuber allerdings deutlich größer. Bei VTubern zeigt sich sogar eine Art Authentizitätsumkehr, bei der eine menschliche Stimme zu Authentizitätskritik führt. Besonders deutlich wird die Authentizitätsumkehr im Voice-Reveal-Video von MOTHERv3 (K020). In dem Video spricht erstmals eine menschliche Stimme statt der bekannten KI-Stimme. Dies löste bei einigen Kommentierenden Verwirrung aus, ob die Stimme zur Figur gehört oder es sich um eine andere Person handelt (K020_V1_S03, K020_V2_S71). Die

menschliche Stimme wurde zwar positiv bewertet (K020_V1_S01) aber auch konkret kritisiert als nicht authentisch: „Cool now I hate you more 1: you're not even an AI, very obviously a human. Please stop claiming you're an AI Vtuber“ (K020_V2_S39). Das zeigt, dass die etablierte KI-Identität der Figur durch die menschliche Stimme nicht klar bestätigt, sondern in Frage gestellt wurde. Auch bei Neuro-sama (K021) sagt ein Kommentar: „It's pretty concerning I'm starting to question whether hes just using a human behind TTS“ (K021_V1_S06). Das zeigt eine Umkehr der Authentizitätskritik, es wird nicht die Menschlichkeit, sondern das KI-Sein hinterfragt (K021_V1_S06). Die Stimme ist bei den VTubern im Material offen als KI gekennzeichnet und eine irrtümliche Menschzuschreibung (AI.3) findet nicht statt. Bei Zentreya (K027) wird die Authentizitätsfrage als Einstieg in das Verständnis der Figur geäußert und meist positiv gerahmt (K027_V2_S12, K027_V2_S14).

4.5.3 Relevanz von KI-Kennzeichnung

Die Akzeptanz der Persona (AI.2) tritt sowohl bei gekennzeichneten als auch nicht gekennzeichneten Videos auf. Beim Kanal Zentreya (K027) erfolgt keine explizite Kennzeichnung der KI-Nutzung. Gleichzeitig lässt sich anhand der Kommentare erkennen, dass die Künstlichkeit der Stimme von den Kommentierenden bereits als bekannt vorausgesetzt wird. Bei transparenter KI-Inszenierung kann eine intensive Auseinandersetzung mit der Grenze zwischen menschlicher und synthetischer Stimme entstehen (K020_V1_S08). RK.0 ist im Persona-Kontext gleichermaßen verteilt auf Kanäle mit und ohne KI-Kennzeichnung. Auffällig ist, dass auch in der Persona-Funktion bei Avatar-Kanälen die Top-Kommentare die Nutzung von KI kaum thematisieren. Bei VTuber-Kanälen wird sie hingegen aktiv als Teil der Figur kommentiert.

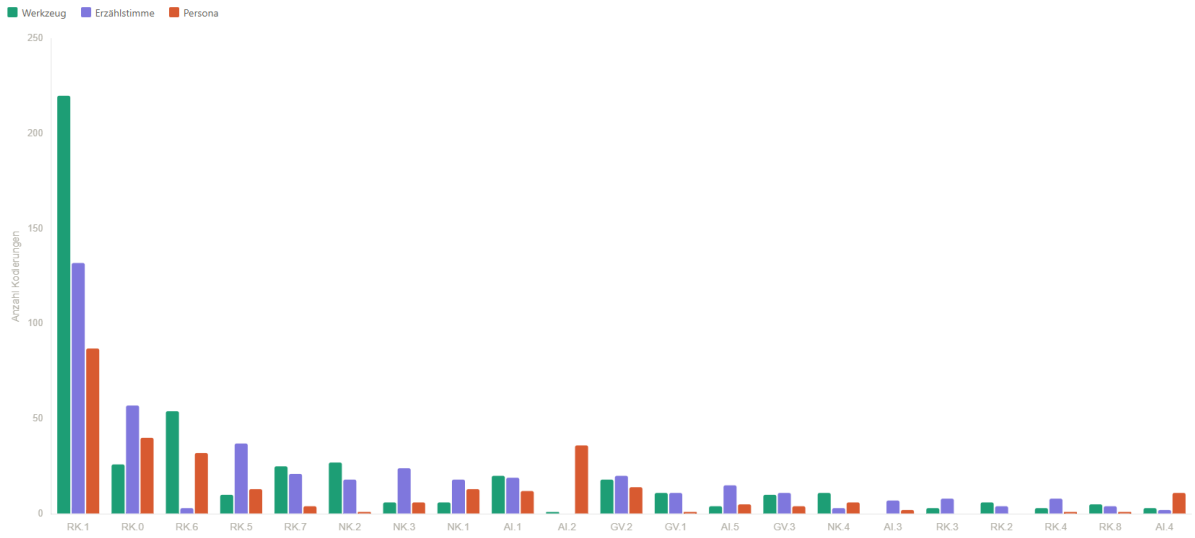
4.6 Funktionsübergreifende Vergleichsmuster

4.6.1 Ähnlichkeiten und Unterschiede

Abbildung 3 zeigt die Verteilung der Unterkategorien-Kodierungen über die drei Stimmfunktionen. Die pauschale Kurzablehnung (RK.1) dominiert in allen Funktionen, fällt jedoch bei Persona deutlich geringer aus.

Abbildung 3

Häufigkeit der Unterkategorien-Kodierungen nach Stimmfunktion



Anmerkung. Dargestellt ist die absolute Anzahl der Kodierungen pro Unterkategorie, aufgeteilt nach Stimmfunktion. Mehrfachkodierung einzelner Kommentare möglich. Die Darstellung dient der visuellen Orientierung; aufgrund des qualitativen Forschungsdesigns sind keine statistischen Schlüsse aus den Häufigkeiten zu ziehen.

Die pauschale Abweisung von KI (RK.1) ist in allen Funktionen dominant vertreten, jedoch deutlich geringer bei der VTuber Untergruppe der Persona-Funktion. Wie Abbildung 3 verdeutlicht, ist RK.1 in der Werkzeug-Funktion am stärksten vertreten, was sich durch die hohe Anzahl an Einzeilern wie „AI slop" erklärt. Besonders präsent im Vergleich zu anderen Funktionen ist hier auch NK.2, allerdings sind die meisten dieser Kommentare auf einen Aussprachefehler in einem Video zurückzuführen (K025_V3_S05). NK.1 ist hingegen in Persona am stärksten vertreten, wobei die Monotonie bei Kanälen der Untergruppe VTuber als Teil der Figur gewertet wird (K021_V1_S02), während sie bei anderen Kanälen funktionsübergreifend eher kritisiert wird (K001_V1_S02). RK.5 trat fast ausschließlich bei der Erzählstimme-Funktion auf. Dabei wurde auf thematische Inkongruenz zwischen Inhalt und KI-Nutzung hingewiesen: „The irony of modern, sloppy, AI narration in a piece on 'made to last' is beyond annoying“ (K013_V1_S13). RK.6 war am stärksten in der Werkzeug-Funktion vertreten und nur vereinzelt in den anderen Funktionen. Parasoziale Enttäuschung nach Erkennung von KI-Nutzung drückten vor allem Kommentare bei der Erzählstimme-Funktion aus. Diese Enttäuschung entsteht in der Persona-Funktion nur dann, wenn anders als erwartet doch ein Mensch hinter einem als KI beschriebenen VTuber steckt. Das Abwendungsverhalten (RK.7) äußert sich auch deutlich stärker in der Erzählstimme- und Werkzeug-Funktion und kaum in der Persona-Funktion. Die Zuschauer akzeptieren bei VTubern die Stimme als Teil der Figur.

4.6.2 Funktionsspezifische und kontextübergreifende Kategorien

Die pauschale Ablehnung von KI-Inhalten (RK.1) ist kontextübergreifend. Ebenso ist der Wunsch nach einer menschlichen Stimme (AI.1) und die Unsicherheit über KI-Nutzung (RK.0)

universell. Auch der fehlende Ausdruck der KI-Stimmen (NK.1) wird in allen Funktionen kommentiert, eine Ausnahme bilden nur Kommentare zur Persona-Unterkategorie, den VTubern, bei denen die Monotonie als positiv bewertet wird.

Die Akzeptanz der Stimme als Teil der Persona bzw. Kanalidentität (AI.2) ist hingegen funktionsspezifisch und tritt nur bei Videos der Persona-Funktion auf. Schwerpunktmäßig funktionsspezifisch sind auch RK.5, RK.3 und AI.5 für Erzählstimme und NK.2 und RK.6 für Werkzeug.

Die positive Klangbewertung (NK.3) findet kontextübergreifend fast ausschließlich bei Gates of Imagination (K010) statt. In ähnlichem Maße findet RK.5 nur bei Industrial Decay (K013) und MonkeyExplains (K019) und RK.3 nur bei Sleepy Time History (K003) statt.

4.6.3 Authentizität im Vergleich

In der Stimmfunktion Werkzeug wirkt Authentizität als Kompetenzfrage. Die KI-Stimme wirkt kompetenzmindernd und führt dazu, dass den Videoinhalten insgesamt weniger vertraut wird. In der Funktion Erzählstimme hingegen wird Authentizität mehr als Vertrauens- und Kulturfrage. Kommentare beschreiben, dass echte menschliche Stimmen mit wahren Emotionen verloren gehen (K008_V1_S22). In der Persona-Funktion, spezifisch bei den VTubern in der Untersuchung, kehrt sich die Wahrnehmung der Authentizität. Die Authentizität der Kanalidentität und Figur wird nicht aufgrund der KI-Stimme hinterfragt, sondern dann, wenn keine KI mehr genutzt wird.

Gemeinsamer Nenner ist mit Ausnahme der VTuber, dass funktionsübergreifend menschliche Stimmen gefordert werden – die Begründung variiert jedoch. Bei der Werkzeug-Funktion geht es um die Kompetenz, bei der Erzählstimme um die Kultur und bei der Persona um die Identität.

4.6.4 Relevanz von KI-Kennzeichnung im Vergleich

Pauschale Ablehnung von KI (RK.1) dominiert sowohl bei gekennzeichneten als auch bei nicht gekennzeichneten Videos in allen drei Stimmfunktionen. Transparenz wird allgemein positiv bewertet (K003_V1_S05), scheint jedoch keinen positiven oder negativen Einfluss auf die Akzeptanz von KI-Nutzung zu haben. Ohne Kennzeichnung tritt im Material häufiger Unsicherheit (RK.0) und Transparenzanforderung (RK.3) auf. Am stärksten zeigt sich die Reduktion von AI.5 und RK.0 bei Kennzeichnung in der Erzählstimme-Funktion. In Persona hat die Kennzeichnung den schwächsten Effekt auf die Akzeptanz der Persona (AI.2).

5 Diskussion

5.1 Gültigkeit und Reichweite der Untersuchung

Die Befunde gelten ausschließlich innerhalb des untersuchten Korpus von 30 Kanälen, 87 englischsprachigen Videos und 2334 Kommentaren und sind nicht auf andere Plattformen, Sprachen oder Medien übertragbar. Aufgrund des explorativen Charakters der Arbeit dienen sie der Hypothesengenerierung für weitere Forschung und explizit nicht der Verallgemeinerung (Stebbins, 2001). Zudem bleibt die Einordnung einzelner Grenzfälle interpretativ, da die Kodierung nur von einer Person durchgeführt wurde. Da YouTube-Kommentare performative öffentliche Äußerungen sind, sind sie nicht eindeutig repräsentativ. Sie dienen lediglich als Grundlage für die explorative Schlussfolgerung zur öffentlichen Aushandlung der Wahrnehmung einer investierten Teilgruppe (Kozinets, 2015).

5.2 Beantwortung der Forschungsfrage

5.2.1 Hauptfrage

Die Hauptfrage dieser Arbeit lautet: „Wie werden KI-generierte Stimmen in YouTube-Kommentaren in Abhängigkeit von ihrer Funktion im Video wahrgenommen und bewertet?“ Innerhalb des Korpus wurde die Nutzung von KI in allen Funktionen überwiegend kritisch wahrgenommen. Funktionsübergreifend dominierte die pauschale Ablehnung von KI (RK.1), während Akzeptanz nur in wenigen Fällen auftrat. Einige Reaktionsmuster traten im Material allerdings auch nur funktionsspezifisch auf. Die Akzeptanz der KI-Stimme als Teil der Kanalidentität oder Persona trat nur in der Persona-Untergruppe der VTuber auf. In dieser Gruppe fand sich zudem wenig KI-Kritik. Dies könnte darauf hindeuten, dass KI-Stimmen, die in dieser Funktion transparent als Bestandteil der Kanalidentität kommuniziert werden, eher akzeptiert werden. Außerdem deutet das Material an, dass KI-Stimmen, besonders in der Erzähler-Funktion davon profitieren, wenn Zuschauer diese fälschlicherweise als menschlich identifizieren, da diese Fehlzuordnung am häufigsten mit positiver Klangbewertung einherging. Es bleibt aber zu beachten, dass der Korpus ein verzerrtes Bild wiedergibt, indem einzelne Kanäle, wie Gates of Imagination (K010), diese Tendenz im Material stark geprägt haben.

5.2.2 Unterfragen

Aus dem Korpus ergeben sich drei Dimensionen der Authentizität nach Stimmfunktion. In der Werkzeugfunktion wirkt Kompetenz ausschlaggebend für die Authentizitätsbewertung, während es bei der Erzählstimme das Vertrauen und bei der Persona die Identität waren. Das lässt vermuten, dass die Formen wahrgenommener Authentizität sich mit der Stimmfunktion ändern.

Es lassen sich keine klaren Konsequenzen für die Medienproduktion ableiten, lediglich Hypothesen formulieren. So deutet das Material darauf hin, dass die Nutzung von KI-Stimmen allgemein kritisch gesehen wird und nur bei klarer Kommunikation als Teil der Kanalidentität in der Persona-Funktion akzeptiert wird.

5.2.3 Vorannahme

Die ursprüngliche Annahme, dass Wahrnehmung und Bewertung synthetischer Stimmen funktionsabhängig variieren, wird im Material nur partiell durch funktionsspezifische Tendenzen gestützt. Dies wird jedoch durch das dominante Grundmuster der pauschalen Ablehnung und kanalspezifische Besonderheiten relativiert. Es kann anhand des Materials weder klar bestätigt noch widerlegt werden, dass Wahrnehmung funktionsabhängig ist.

5.3 Interpretation im theoretischen Rahmen

5.3.1 Reaktion auf KI-Einsatz

Die Ergebnisse decken sich mit den Befunden von Fan & Liu (2025), dass Stimmen anders bewertet werden, wenn sie als KI gekennzeichnet sind. Kennzeichnung muss im Rahmen dieser Arbeit dafür allerdings weiter definiert werden als bisher. Im Material zeigt sich: Die Stimme wird fast ausschließlich positiv bewertet, wenn Kommentierende sie fälschlicherweise einem Menschen zuschreiben, während sie bei Erkennung deutlich negativer bewertet wurde. Die Identifikation der KI-Nutzung fand im Material fast unabhängig davon statt, ob die Kanäle es selbst kennzeichneten oder nicht. Demnach ist es möglich, dass der Label-Effekt auch auf YouTube eintritt, das Label allerdings nur begrenzt von den Kanälen selbst beeinflusst werden kann. Außerdem deuten die Ergebnisse an, dass der Plattformkontext den Label-Effekt im Vergleich zu Laboruntersuchungen verstärkt. Eine Ausnahme im Material, in dem der Label-Effekt möglicherweise zur positiven Stimmbewertung beigetragen hat, ist der Kanal Gates of Imagination (K010), der den KI-Stimmen Namen zuweist und sie somit indirekt als Menschen kennzeichnet.

5.3.2 Natürlichkeit und Klang

Hinterleitner (2017) beschreibt Natürlichkeit als stärksten Prädiktor für die Bewertung synthetischer Stimmen. Wenn die pauschale Ablehnung aufgrund der KI-Herkunft einer Stimme ausgenommen wird, fällt im Material in der Werkzeugfunktion jedoch vorwiegend die nicht vorhandene Störungsfreiheit (NK.2) auf. Dies kann allerdings daran liegen, dass Natürlichkeit vorausgesetzt und nicht kommentiert wird. Das Material umfasst auch nur Stimmen gleicher Natürlichkeit, wodurch diese Bewertungsebene auch innerhalb des Materials nicht klar eingestuft werden kann.

Der Korpus unterstützt auch die Aussage von Larrouy-Maestri et al. (2025), dass spontan produzierte Sprache emotionale Zustände ausdrücken kann, die KI nicht replizieren kann. Dies zeigt sich durch das häufige Auftreten der Kategorie NK.1 und durch Kommentare, wie diesen: „Rest in peace, the days when actual humans used their actual human voices to convey real emotion“ (K008_V1_S22). Das zeigt, dass auf YouTube möglicherweise in allen drei Stimmfunktionen, besonders in narrativen Kontexten, aktuelle KI-Stimmen die Erwartungen an emotionale Prosodie noch nicht vollständig erfüllen können. Im Widerspruch dazu steht der Kanal Gates of Imagination (K010), bei dem die KI-Stimmen positiv bewertet werden. In diesem Fall scheinen die Anforderungen an emotionale Prosodie erfüllt. Es ist also möglich, dass die Qualität der Stimmen in solchen Fällen ausreichend ist. Faktoren, die in diesem Fall positiv zur Akzeptanz beitragen könnten und in dieser Untersuchung nicht überprüft werden können, sind demografische Angaben der Kommentierenden, mögliche Zensur von negativen Kommentaren und die Namensgebung der Stimmen, die einen Menschen hinter der Stimme impliziert.

5.3.3 Glaubwürdigkeit und Kontext

Bezogen auf die Kategorie Glaubwürdigkeit und Vertrauen zeigt sich im Material, dass Kontext eine zentrale Rolle spielt. Die Akzeptanz KI-generierter Stimmen weist funktionsspezifische Unterschiede auf und ist demnach auch auf YouTube möglicherweise kontextabhängig. Das entspricht den Befunden von Schreibelmayr & Mara (2022). Im Material zeigte sich auch, dass besonders, wenn der Kontext als unpassend angesehen (RK.5) wurde, die Bewertung der Stimme negativer ausfiel. Das deutet darauf hin, dass auf YouTube der Kontext der Stimmnutzung eine wichtige Rolle für die Akzeptanz spielt.

5.3.4 Authentizität und Identität

Im Korpus zeigt sich die Forderung nach menschlichem Sprecher (AI.1) bei allen Stimmfunktionen und deutet an, dass auch bei KI-Stimmen das wichtigste Authentizitätsmerkmal eine klare Präsenz einer echten Persönlichkeit ist, wie Riboni (2020) beschreibt. Auch Bruder et al. (2025) werden im Material dadurch bestätigt, dass die Bewertung der KI-Stimmen überwiegend negativ ausfällt und vermehrt die direkte Forderung nach einer menschlichen Stimme (AI.1) geäußert wird.

Im Material zeigt sich auch, dass parasoziale Beziehungen, wie sie von Hartmann & Goldhoorn (2011) beschrieben wurden, möglicherweise auch durch KI-Stimmen ausgelöst werden können. Besonders in Persona-Funktion, wo die KI-Stimme als Teil der Identität akzeptiert wird, zeigt sich, dass eine menschliche Stimme negative Auswirkungen auf die parasoziale Beziehung haben kann, wenn sie zuvor Teil der Identität war. In den anderen Funktionen zeigt sich lediglich die Enttäuschung nach Erkennung der KI-Stimme. Das könnte bedeuten, dass die Stimme in Persona-Funktion bei offener Kommunikation als Teil der Identität parasoziale

Bindung erzeugen kann, während das in anderen Funktionen nur möglich ist, solange die Stimme nicht als künstlich erkannt wird.

In Widerspruch zu Sundar & Nass (2000), die zeigten, dass Nutzer nicht den Programmierer, sondern den Computer selbst als Informationsquelle wahrnehmen, wurde im Material häufig getrennt die Person hinter dem Kanal kritisiert und nicht die Stimme. Der VTuber-Befund, dass eine KI-Stimme in einigen Fällen als authentisch und eine menschliche als nicht authentisch wahrgenommen wird, deutet darauf hin, dass bestehende Forschungsergebnisse, bei denen menschliche Stimmen konstant als authentischer gewertet werden (Chen et al., 2025; Kühne et al., 2020), nicht direkt auf den YouTube-Kontext übertragbar sind.

5.4 Konsequenzen für Medienproduktion und Gestaltung

Für Medienproduzierende lassen sich auf Basis der Untersuchung folgende vorläufige Orientierungspunkte in Form von Hypothesen festhalten:

Bei der Stimme als Persona deutet das Material an, dass offene Kommunikation der KI-Nutzung als Teil der Kanalidentität ausschlaggebend für die Akzeptanz ist. Eine konsistente Stimme über Videos hinweg scheint dabei relevanter zu sein als Natürlichkeit und Klang der Stimme.

Bei der Stimme als Erzählstimme legt das Material nahe, dass Passung und thematische Kohärenz zwischen Inhalt und KI-Einsatz besonders wichtig sind. Videos, die sich kritisch mit Technik oder menschlicher Arbeit befassen **und** klanglich als unpassend wahrgenommene Stimmen nutzen, scheinen besonders anfällig für Kritik zu sein.

Bei der Stimme als Werkzeug deutet das Material darauf hin, dass technische Qualität, besonders korrekte Aussprache und angemessenes Tempo, die höchste Relevanz haben. Die Frage, ob KI genutzt wird scheint hier weniger entscheidend als Störungsfreiheit.

Übergreifend deutet das Material an, dass die Entscheidung, KI-Stimmen auf YouTube zu nutzen, bewusst getroffen werden sollte mit Blick auf Stimmfunktion, Thema und Kennzeichnung. Kommentare deuten darauf hin, dass die Nutzung menschlicher Stimmen funktionsübergreifend die beste Möglichkeit ist, um die Glaubwürdigkeit und Authentizität des Kanals zu bewahren.

5.5 Methodische Grenzen und alternative Deutungen

Aufgrund des explorativen Forschungsdesigns lassen sich keine gesicherten Handlungsempfehlungen aus der Untersuchung ableiten, sondern nur Hypothesen, die weiter überprüft werden müssen. Auch Kausalaussagen und Repräsentativität sind nicht möglich, ebenso wenig Aussagen, die über die Plattform, Sprache und das Medium hinausgehen.

5.5.1 Unsicherheit der Deutungen

Die Funktion als Ursache ist im Material in einigen Fällen nicht belegbar, da einzelne Kanäle die Masse an Kommentaren einer Kategorie ausmachen. Die Deutung, dass in der Werkzeug-Funktion besonders auf korrekte Aussprache und Verständlichkeit geachtet wird ist fast ausschließlich dem Kanal BoNa (K025) zuzuordnen, da dieser den Großteil der NK.2-Kommentare ausmacht. Die reinen Zahlen sollten keine Basis für Deutungen sein.

5.5.2 Grenzen des Materials

Durch die Schweigespirale nach Noelle-Neumann (1974) können bereits vertretene Meinungen verstärkt und andere geschwächt werden, die pauschale Abneigung gegen KI kann dadurch also stärker sichtbar sein als sie tatsächlich vorhanden ist. Dieses Phänomen kann durch Bot-Kommentare verstärkt werden, die selbst eine weitere Grenze bilden. Auch demografische Daten können nicht erfasst werden und zur Verzerrung der Ergebnisse beitragen. Auch die zuvor angesprochene Zensur bildet eine Variable des Materials, die nicht nachvollziehbar ist.

5.5.3 Andere Einflussfaktoren für Kommentarreaktionen

Ein wichtiger Faktor abseits der Funktionen war im Material auch der Inhalt. So trat besonders viel Kritik bei Inkongruenz zwischen Videoinhalt und KI-Nutzung auf, was möglicherweise bedeutet, dass Videoinhalte ein wichtiger Faktor in der Bewertung von KI-Stimmen-Nutzung sind. Videos, die abseits der Stimme noch andere erkennbare Formen der KI-Nutzung zeigten, wurden stärker negativ bewertet als solche, die nur die Stimme nutzen. Dies lässt sich aus den Daten jedoch nicht eindeutig belegen.

5.6 Hypothesen

Aus den Befunden lassen sich folgende Hypothesen für KI-Stimmen auf YouTube ableiten, die in weiterer Forschung überprüft werden müssen:

H1: KI-Stimmen werden in der Persona-Funktion eher akzeptiert, wenn sie offen als Teil der Kanalidentität kommuniziert werden.

H2: Thematische Inkongruenz zwischen Videoinhalt und KI-Einsatz verstärkt negative Reaktionen unabhängig von der Stimmqualität.

H3: Irrtümliche Zuschreibung einer KI-Stimme als menschlich beeinflusst die Klangbewertung positiv.

H4: In funktionalen Kontexten wirkt sich die fehlerhafte Aussprache stärker negativ auf die Wahrnehmung aus.

H5: Fehlende Kennzeichnung erhöht Unsicherheit und Transparenzforderungen, aber nicht zwangsläufig die Akzeptanz.

5.7 Konklusion und Ausblick

Die vorliegende Arbeit untersuchte explorativ, wie KI-generierte Stimmen in YouTube-Kommentaren in Abhängigkeit von ihrer Funktion im Video wahrgenommen und bewertet werden. Die Untersuchung deutet auf ein plattformweites Grundmuster pauschaler Ablehnung und funktionsspezifische Tendenzen hin: technische Klangkritik konzentriert sich im Werkzeug-Kontext, inhaltliche Inkongruenz bei Erzählstimmen und Akzeptanz der Stimme als Teil der Identität im Persona-Kontext. Diese Tendenzen sind Hypothesen und keine gesicherten Befunde, da das explorative Design keine Kausalaussagen erlaubt.

Für die Medienproduktion lässt sich als Orientierung festhalten, dass die Untersuchung darauf hindeutet, dass die Notwendigkeit der Nutzung einer KI-Stimme auf YouTube hinterfragt werden sollte, um Kritik zu vermeiden. Auch indiziert sie, dass die Nutzung offengelegt, als Teil der Kanalidentität kommuniziert und thematisch passend eingesetzt werden sollte. Um diese Hypothesen zu prüfen, wäre ein kontrolliertes Design notwendig, das möglichst viele Variablen dieser Untersuchung einschränkt. Dazu gehört die Nutzung der gleichen Stimme, die gleiche Art der Kennzeichnung und das Sammeln demografischer Daten. Dafür könnten beispielsweise Umfragen durchgeführt werden, die mehr Informationen der Teilnehmenden abfragen, die aus Kommentaren nicht hervorgehen. Konkret könnten experimentelle Studien durchgeführt werden, bei denen dieselbe Stimme in verschiedenen Funktionen bewertet wird, um den Einfluss der Funktion isoliert zu messen. Eine quantitative Inhaltsanalyse eines größeren Kommentarkorpus könnte die gefundenen Tendenzen der vorliegenden Untersuchung statistisch prüfen. Variablen wie Kennzeichnung, Stimmqualität und thematische Passung sollten dabei kontrolliert werden. Auch demografische Daten der Kommentierenden, etwa Alter, Geschlecht und Vorwissen über KI, sollten erfasst werden, da sie basierend auf bestehender Forschung nachweislich die Wahrnehmung synthetischer Stimmen beeinflussen. Es bleibt offen, ob das Grundmuster der pauschalen Ablehnung von KI-Stimmen stabil bleibt, wenn sie zur Normalität in der Produktion von YouTube-Videos werden.

Synthetische Stimmen sind in der Medienproduktion schon lange keine Seltenheit mehr und wie mit ihnen gestalterisch umgegangen wird, ist eine Frage, die Medienproduzierende zunehmend beschäftigen wird und bisher empirisch kaum fundiert ist. Diese Arbeit ist ein erster Schritt in diese Richtung.

Literaturverzeichnis

AIVoiceDetector.com. (o. J.). *AI Voice Detector*. Abgerufen 2. April 2026, von https://aivoice-detector.com/

Alali, A., & Theodorakopoulos, G. (2025). Partial Fake Speech Attacks in the Real World Using Deepfake Audio. *Journal of Cybersecurity and Privacy*, *5*(1), 6. https://doi.org/10.3390/jcp5010006

Anthropic. (2025). Advancing Claude for Financial Services. *Anthropic*. https://www.anthropic.com/news/advancing-claude-for-financial-services

Azzuni, H., & Saddik, A. E. (2025). *Voice Cloning: Comprehensive Survey* (arXiv:2505.00579). arXiv. https://doi.org/10.48550/arXiv.2505.00579

Baird, A., Parada-Cabaleiro, E., Hantke, S., Burkhardt, F., Cummins, N., & Schuller, B. (2018). The Perception and Analysis of the Likeability and Human Likeness of Synthesized Speech. *Interspeech 2018*, 2863–2867. https://doi.org/10.21437/Interspeech.2018-1093

Bruder, C., Breda, P., & Larrouy-Maestri, P. (2025). Attractive synthetic voices. *Computers in Human Behavior: Artificial Humans*, *6*, 100211. https://doi.org/10.1016/j.chbah.2025.100211

Bullingham, L., & Vasconcelos, A. C. (2013). 'The presentation of self in the online world': Goffman and the study of online identities. *Journal of Information Science*, *39*(1), 101–112. https://doi.org/10.1177/0165551512470051

Çalli, L., & Alma Çalli, B. (2025). Recoding Reality: A Case Study of YouTube Reactions to Generative AI Videos. *Systems*, *13*(10), 925. https://doi.org/10.3390/systems13100925

Cambre, J., & Kulkarni, C. (2019). One Voice Fits All? Social Implications and Research Challenges of Designing Voices for Smart Devices. *Proc. ACM Hum.-Comput. Interact.*, *3*(CSCW), 223:1-223:19. https://doi.org/10.1145/3359325

Chen, W., Pell, M. D., & Jiang, X. (2025). *Does Speech Prosody Shape Social Perception Equally for AI and Human Voices? A 16-Dimension Rating Study*. Social Sciences. https://doi.org/10.20944/preprints202510.1492.v1

Chion, M. (1999). *The voice in cinema* (Nachdr.). Columbia University Press.

Cooper, E., Huang, W.-C., Tsao, Y., Wang, H.-M., Toda, T., & Yamagishi, J. (2024). A review on subjective and objective evaluation of synthetic speech. *Acoustical Science and Technology*, *45*(4), 161–183. https://doi.org/10.1250/ast.e24.12

Dynel, M. (2014). Participation framework underlying YouTube interaction. *Journal of Pragmatics*, *73*, 37–52. https://doi.org/10.1016/j.pragma.2014.04.001

ElevenLabs. (o. J.). *AI Speech Classifier*. Abgerufen 2. April 2026, von https://elevenlabs.io/de/ai-speech-classifier

ElevenLabs. (2024, Januar 22). *ElevenLabs veröffentlicht neue KI-Produkte und sammelt 80 Mio. USD in Serie B*. ElevenLabs. https://elevenlabs.io/de/blog/series-b

Fan, G., & Liu, D. (2025). *When Machines Speak with Feeling: Investigating Emotional Prosody, Authenticity, and Trust in AI vs. Human Voices*. (Volume 47). https://escholarship.org/uc/item/8vr8s6h8

FindAIVoice.com. (o. J.). *Find AI Voice*. Abgerufen 2. April 2026, von https://findaivoice.com/

Gambino, A., Fox, J., & Ratan, R. (2020). Building a Stronger CASA: Extending the Computers Are Social Actors Paradigm. *Human-Machine Communication*, *1*, 71–86. https://doi.org/10.30658/hmc.1.5

Goffman, E. (1956). *The Presentation of Self in Everyday Life* (1. Aufl.). Social Sciences Research Centre, University of Edinburgh.

Gong, C. (2023). AI voices reduce cognitive activity? A psychophysiological study of the media effect of AI and human newscasts in Chinese journalism. *Frontiers in Psychology*, *14*, 1243078. https://doi.org/10.3389/fpsyg.2023.1243078

Hartmann, T., & Goldhoorn, C. (2011). Horton and Wohl Revisited: Exploring Viewers' Experience of Parasocial Interaction. *Journal of Communication*, *61*(6), 1104–1121. https://doi.org/10.1111/j.1460-2466.2011.01595.x

Hinterleitner, F. (2017). *Quality of Synthetic Speech*. Springer Singapore. https://doi.org/10.1007/978-981-10-3734-4

Hogan, B. (2010). The Presentation of Self in the Age of Social Media: Distinguishing Performances and Exhibitions Online. *Bulletin of Science, Technology & Society*, *30*(6), 377–386. https://doi.org/10.1177/0270467610385893

Honeycutt, J. M., & Bryan, S. P. (2010). *Scripts and Communication for Relationships*. https://www.peterlang.com/document/1143731 (Ursprünglich erschienen Peter Lang Verlag)

Horton, D., & Wohl, R. R. (1956). Mass Communication and Para-Social Interaction: Observations on Intimacy at a Distance. *Psychiatry*, *19*(3), 215–229. https://doi.org/10.1080/00332747.1956.11023049

Im, H., Sung, B., Lee, G., & Xian Kok, K. Q. (2023). Let voice assistants sound like a machine: Voice and task type effects on perceived fluency, competence, and consumer attitude. *Computers in Human Behavior*, *145*, 107791. https://doi.org/10.1016/j.chb.2023.107791

Jost, C. (2024). „Thank You for Sharing this Fantastic Performance": Meaning and Form in Transmedia Persona Construction of YouTube Drummers. *Persona Studies*, *10*(1), 85–98. https://doi.org/10.21153/psj2024vol10no1art1870

Kang, E. J., Kim, H., Kim, H., Fussell, S. R., & Kim, J. (2025). Can Fans Build Parasocial Relationships through Idols' Simulated Voice Messages?: A Study of AI Private Call Users' Perceptions, Cognitions, and Behaviors. *Proceedings of the ACM on Human-Computer Interaction*, *9*(2), 1–31. https://doi.org/10.1145/3711111

Khan, M. L. (2017). Social media engagement: What motivates user participation and consumption on YouTube? *Computers in Human Behavior*, *66*, 236–247. https://doi.org/10.1016/j.chb.2016.09.024

Kozinets, R. V. (2015). *Netnography: Redefined* (2.). SAGE.

Kozloff, S. (with American Council of Learned Societies). (1988). *Invisible storytellers: Voice-over narration in American fiction film*. University of California Press.

Kuckartz, U. (2016). *Qualitative Inhaltsanalyse. Methoden, Praxis, Computerunterstützung* (3., überarbeitete Aufl). Beltz.

Kühne, K., Fischer, M. H., & Zhou, Y. (2020). The Human Takes It All: Humanlike Synthesized Voices Are Perceived as Less Eerie and More Likable. Evidence From a Subjective Ratings Study. *Frontiers in Neurorobotics*, *14*, 593732. https://doi.org/10.3389/fnbot.2020.593732

Larrouy-Maestri, P., Poeppel, D., & Pell, M. D. (2025). The Sound of Emotional Prosody: Nearly 3 Decades of Research and Future Directions. *Perspectives on Psychological Science*, *20*(4), 623–638. https://doi.org/10.1177/17456916231217722

Maltezou-Papastylianou, C., Scherer, R., & Paulmann, S. (2025a). Evaluating trustworthiness across ethnically diverse human and commercial synthesised voices: A comparative study. *Computers in Human Behavior Reports*, *19*, 100762. https://doi.org/10.1016/j.chbr.2025.100762

Maltezou-Papastylianou, C., Scherer, R., & Paulmann, S. (2025b). How do voice acoustics affect the perceived trustworthiness of a speaker? A systematic review. *Frontiers in Psychology*, *16*, 1495456. https://doi.org/10.3389/fpsyg.2025.1495456

Marwick, A. E., & Boyd, D. (2011). I tweet honestly, I tweet passionately: Twitter users, context collapse, and the imagined audience. *New Media & Society*, *13*(1), 114–133. https://doi.org/10.1177/1461444810365313

Metricool. (2025, August 19). *Für welche Social-Media-Aufgaben verwenden sie KI?* https://de.statista.com/statistik/daten/studie/1478575/umfrage/einsatz-von-kuenstlicher-intelligenz-in-sozialen-medien-durch-unternehmen-weltweit/

Möller, A. M., Baumgartner, S. E., Kühne, R., & Peter, J. (2021). Sharing the Fun? How Social Information Affects Viewers' Video Enjoyment and Video Evaluations. *Human Communication Research*, *47*(1), 25–48. https://doi.org/10.1093/hcr/hqaa013

Möller, A. M., Kühne, R., Baumgartner, S. E., & Peter, J. (2019). Exploring User Responses to Entertainment and Political Videos: An Automated Content Analysis of YouTube. *Social Science Computer Review*, *37*(4), 510–528. https://doi.org/10.1177/0894439318779336

Möller, A. M., Vermeer, S. A. M., & Baumgartner, S. E. (2024). Cutting Through the Comment Chaos: A Supervised Machine Learning Approach to Identifying Relevant YouTube Comments. *Social Science Computer Review*, *42*(1), 162–185. https://doi.org/10.1177/08944393231173895

Mori, M., MacDorman, K., & Kageki, N. (2012). The Uncanny Valley [From the Field]. *IEEE Robotics & Automation Magazine*, *19*(2), 98–100. https://doi.org/10.1109/MRA.2012.2192811

Nass, C., & Moon, Y. (2000). Machines and Mindlessness: Social Responses to Computers. *Journal of Social Issues*, *56*(1), 81–103. https://doi.org/10.1111/0022-4537.00153

Ning, Y., He, S., Wu, Z., Xing, C., & Zhang, L.-J. (2019). A Review of Deep Learning Based Speech Synthesis. *Applied Sciences*, *9*(19), 4050. https://doi.org/10.3390/app9194050

Noelle-Neumann, E. (1974). The Spiral of Silence a Theory of Public Opinion. *Journal of Communication*, *24*(2), 43–51. https://doi.org/10.1111/j.1460-2466.1974.tb00367.x

Noufi, C., May, L., & Berger, J. (2025). A model of vocal persona: Context, perception, production. *Frontiers in Computer Science*, *7*, 1575296. https://doi.org/10.3389/fcomp.2025.1575296

Nussbaum, C., Frühholz, S., & Schweinberger, S. R. (2025). Understanding voice naturalness. *Trends in Cognitive Sciences*, *29*(5), 467–480. https://doi.org/10.1016/j.tics.2025.01.010

Orynbay, L., Razakhova, B., Peer, P., Meden, B., & Emeršič, Ž. (2024). Recent Advances in Synthesis and Interaction of Speech, Text, and Vision. *Electronics*, *13*(9), 1726. https://doi.org/10.3390/electronics13091726

Poché, E., Jha, N., Williams, G., Staten, J., Vesper, M., & Mahmoud, A. (2017). Analyzing User Comments on YouTube Coding Tutorial Videos. *2017 IEEE/ACM 25th International Conference on Program Comprehension (ICPC)*, 196–206. https://doi.org/10.1109/ICPC.2017.26

Reeves, B., & Nass, C. (1996). The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Pla. *Bibliovault OAI Repository, the University of Chicago Press*.

Riboni, G. (2020). *Discourses of Authenticity on YouTube: From the Personal to the Professional*. LED Edizioni Universitarie di Lettere Economia Diritto. https://www.ledonline.it/public/files/journals/9/931-8/authenticity-youtube.pdf

Roesler, E., Manzey, D., & Onnasch, L. (2021). A meta-analysis on the effectiveness of anthropomorphism in human-robot interaction. *Science Robotics*, *6*(58), eabj5425. https://doi.org/10.1126/scirobotics.abj5425

Romportl, J. (2014). Speech Synthesis and Uncanny Valley. In P. Sojka, A. Horák, I. Kopeček, & K. Pala (Hrsg.), *Text, Speech and Dialogue* (Bd. 8655, S. 595–602). Springer International Publishing. https://doi.org/10.1007/978-3-319-10816-2_72

Schank, R. C., & Abelson, R. P. (1977). *Scripts, plans, goals and understanding: An inquiry into human knowledge structures* (S. 248). Lawrence Erlbaum.

Schreibelmayr, S., & Mara, M. (2022). Robot Voices in Daily Life: Vocal Human-Likeness and Application Context as Determinants of User Acceptance. *Frontiers in Psychology*, *13*, 787499. https://doi.org/10.3389/fpsyg.2022.787499

Shiramizu, V. K. M., Lee, A. J., Altenburg, D., Feinberg, D. R., & Jones, B. C. (2022). The role of valence, dominance, and pitch in perceptions of artificial intelligence (AI) conversational agents' voices. *Scientific Reports*, *12*(1), 22479. https://doi.org/10.1038/s41598-022-27124-8

Stebbins, R. (2001). *Exploratory Research in the Social Sciences*. SAGE Publications, Inc. https://doi.org/10.4135/9781412984249

Stern, S. E., Mullennix, J. W., & Yaroslavsky, I. (2006). Persuasion and social perception of human vs. Synthetic voice across person as source and computer as source conditions. *International Journal of Human-Computer Studies*, *64*(1), 43–52. https://doi.org/10.1016/j.ijhcs.2005.07.002

Sui, W., Sui, A., & Rhodes, R. E. (2022). What to watch: Practical considerations and strategies for using YouTube for research. *DIGITAL HEALTH*, *8*, 205520762211237. https://doi.org/10.1177/20552076221123707

Sundar, S. S., & Nass, C. (2000). Source Orientation in Human-Computer Interaction: Programmer, Networker, or Independent Social Actor. *Communication Research*, *27*(6), 683–703. https://doi.org/10.1177/009365000027006001

Taake, K. (2009). *A Comparison of Natural and Synthetic Speech: With and Without Simultaneous Reading*. Washington University.

Thelwall, M. (2018). Social media analytics for YouTube comments: Potential and limitations. *International Journal of Social Research Methodology*, *21*(3), 303–316. https://doi.org/10.1080/13645579.2017.1381821

TruthScan. (o. J.). *AI Voice Detector*. Abgerufen 2. April 2026, von https://truthscan.com/de/ai-voice-detector

Undetectable AI. (o. J.). *AI Voice Detector*. Abgerufen 2. April 2026, von https://undetectable.ai/de/ai-voice-detector

Voorveld, H., Panteli, A., Schirris, Y., Ischen, C., Kanoulas, E., & Lentz, T. (2025). Examining the persuasiveness of text and voice agents: Prosody aligned with information structure increases human-likeness, perceived personalisation and brand attitude. *Behaviour & Information Technology*, *44*(12), 2913–2928. https://doi.org/10.1080/0144929X.2024.2420871

We Are Social, DataReportal, & Meltwater. (2025, Oktober 15). *Ranking der größten Social Networks und Messenger nach der Anzahl der Nutzer im Oktober 2025 (in Millionen)*. https://de.statista.com/statistik/daten/studie/181086/umfrage/die-weltweit-groessten-social-networks-nach-anzahl-der-user/

WeCreate. (2025, September 19). *Akzeptanz von KI-generierten Inhalten auf Social Media unter der Generation Z und Generation Y in Deutschland im Jahr 2025*. https://de.statista.com/statistik/daten/studie/1624930/umfrage/akzeptanz-von-ki-content-auf-social-media/

Xie, T., Rong, Y., Zhang, P., Wang, W., & Liu, L. (2025). *Towards Controllable Speech Synthesis in the Era of Large Language Models: A Systematic Survey* (arXiv:2412.06602). arXiv. https://doi.org/10.48550/arXiv.2412.06602

Anhang A – Digitaler Anhang

Die folgenden Dateien werden als ergänzende Materialien zusammen mit dieser Arbeit eingereicht. Sie dienen der Nachvollziehbarkeit der Datenerhebung und Kodierung. Eine kompakte Übersicht des Kategoriensystem bietet Tabelle 2 im Haupttext.

Dateiname	Inhalt
Cornelsen_Videokorpus.xlsx	Vollständiger Videokorpus mit 87 Videos von 30 Kanälen, einschließlich Metadaten (Titel, URL, Veröffentlichungsdatum, Aufrufzahl), Stimmfunktionszuordnung, KI-Kennzeichnung und Zuordnungsprotokollen
Cornelsen_Kommentarkorpus.xlsx	Vollständiger Kommentarkorpus mit 1044 kodierten relevanten Kommentaren und 862 Top-Kommentaren, einschließlich Kategorienzuordnung, Valenzkodierung und Kommentar-IDs
Cornelsen_Kodierbuch.xlsx	Vollständiges Kodierbuch mit Definitionen aller Haupt- und Unterkategorien, Ankerbeispielen, Abgrenzungsregeln und Valenzhinweisen

Anhang B – Erhebungstools

Die folgenden Web-Apps wurden vom Verfasser mithilfe von Google AI Studio entwickelt und dienten ausschließlich der technischen Datenerhebung. Die Interpretation und Kodierung der Daten erfolgten manuell.

Name	Zweck	Technische Basis	Ausgabeformat	Verfügbarkeit
Top Video Finder	Identifikation der meistgeklickten Videos je Kanal; optionale Ausgabe der Top-Kommentare	Google AI Studio, YouTube Data API v3	TSV	Link
Top Comment Finder	Strukturierter Export der Top-Kommentare inkl. aller relevanter Metadaten je Video	Google AI Studio, YouTube Data API v3	CSV	Link
Video Data Extractor	Export relevanter Videodaten (Titel, Aufrufe, Datum etc.) auf Basis einzelner Video-Links	Google AI Studio, YouTube Data API v3	CSV	Link
AI-Voice Comment Finder	Filterung aller Kommentare eines Videos nach den Stichworten „ai“ und „voice“	Google AI Studio, YouTube Data API v3	CSV	Link

Anmerkung. Alle Apps nutzen die YouTube Data API v3 zum Zugriff auf öffentlich verfügbare Video- und Kommentardaten. Unter Verfügbarkeit kann über einen Hyperlink die jeweilige Web-App geöffnet werden. Die ausgeschriebenen Links sind auf Anfrage beim Verfasser erhältlich.

Selbstständigkeitserklärung

Hiermit erkläre ich, dass ich die vorliegende Arbeit mit dem Titel „KI-generierte Stimmen in unterschiedlichen Nutzungskontexten auf YouTube: Eine explorative, vergleichende Analyse von YouTube-Kommentaren“ selbstständig, ohne unerlaubte fremde Hilfe verfasst und keine anderen als die angegebenen Hilfsmittel benutzt habe. Die Stellen der Arbeit, die dem Wortlaut oder dem Sinn nach anderen Werken (dazu zählen auch Internetquellen) entnommen sind, wurden unter Angabe der Quelle kenntlich gemacht.

Erklärung zur Nutzung von KI-Werkzeugen

Im Rahmen dieser Arbeit wurden KI-basierte Werkzeuge unterstützend eingesetzt. Die Nutzung erfolgte im Sinne der KI-Leitlinie der TH OWL transparent, kritisch-reflektiert und unter Wahrung guter wissenschaftlicher Praxis. Die inhaltliche Verantwortung für diese Arbeit liegt vollständig bei mir.

Eingesetzte Werkzeuge:

ResearchRabbit, **Elicit** und **Claude** zur KI-gestützten Literaturrecherche

Claude (Anthropic; u. a. Sonnet 4.5, Sonnet 4.6, Opus 4.6, Opus 4.7), **ChatGPT** (OpenAI; GPT-5) und **Gemini** (Google; Gemini 3 Pro) für Formulierungshilfe, Code-Unterstützung und Kanalrecherche

Konkrete Einsatzbereiche:

Literaturrecherche: Unterstützung bei der Identifikation potenziell relevanter Quellen über ResearchRabbit, Elicit und Claude. Alle Treffer wurden eigenständig im Original gesichtet, auf Seriosität geprüft und ausgewertet.

Recherche relevanter Kanäle: Unterstützung bei der Vorauswahl der in Kapitel 3.4 beschriebenen Kanäle. Die abschließende Auswahl und Bewertung erfolgten eigenständig nach den dort dargelegten Kriterien.

Formulierungshilfe: Punktuelle sprachliche Überarbeitung/ Formulierungshilfe einzelner Textstellen. Inhaltliche Aussagen und Argumentation stammen ausschließlich von mir.

Datenaufbereitung: Unterstützung bei der tabellarischen und farblichen Strukturierung der Rohdaten in Excel mittels Claude Add-In.

Entwicklung der Anwendungen: Unterstützung bei der Implementierung der in Kapitel 3.4.1 beschriebenen Apps (Erstellung nach Anweisung).

Nicht eingesetzt wurde KI zur Generierung inhaltlicher Passagen, zur Interpretation von Ergebnissen oder zur Formulierung von Schlussfolgerungen.

Ort, Datum: Detmold den 11.05.2026

Unterschrift: