

Blogbeitrag vom 25. März 2024

INTELLIGIBLE TEXTUREN

Welche Rolle kann ChatGPT bei der Aufsatzbewertung spielen?

Jan Georg Schneider

Der vorliegende Beitrag behandelt die Frage, welche Rolle ChatGPT bei der Aufsatzbewertung spielen kann. Anhand verschiedener Textausgaben von GPT3.5 und 4 wird im Einzelnen überprüft, wie stark der Bezug zwischen maschinell generierter ‚Notenbegründung‘ und zu bewertendem Aufsatz ist. Insbesondere wird getestet, ob der Chatbot Fehler verschiedener Art (sprachlich-stilistische, logische, sachlich-referentielle) identifizieren und kategorial unterscheiden kann. Die Untersuchung dient dazu, Folgendes herauszufinden: Kann ChatGPT unter Umständen eine Notenbegründung „leistungsgleich ersetzen“ (Janich) oder zumindest dabei helfen, eine solche Begründung zu formulieren? Die dahintersteckende größere Frage, die sich auch auf andere Textsorten und Sprachspiele (im Sinne Wittgensteins) übertragen ließe, lautet: Reicht die massenhafte syntaktische Exemplifikation von Wörtern in realen Kontexten, mit der ChatGPT trainiert wird, unter Umständen (irgendwann) aus, um die mangelnde Referentialität auszugleichen?



| VK:KIWA

1 Einleitung – von der Notenvorhersage zur Notenbegründung¹

Die Funktionsweise automatisierter Aufsatzbewertung wurde von Schneider & Zweig (2022 und 2023) beschrieben und aus semiotisch-sprachphilosophischer Perspektive analysiert. Am Beispiel des US-amerikanischen E-Raters wurde dort ausgeführt, dass solche technischen Systeme, auch wenn sie vorgeblich einen Text bewerten (*to rate*), in Wirklichkeit natürlich nichts anderes als eine Notenvorhersage abgeben. Wenn man davon ausgeht, dass Maschinen generell keine Handlungen im engeren Sinne vollziehen, könnte der Anspruch maximal darin bestehen, eine menschliche Handlung im Sinne von Peter Janich „leistungsgleich“ zu „ersetzen“ (Janich 2015, 302 et passim, Janich 2006, 161–165). Hierzu müssten in unserem Falle allerdings die Gelingensbedingungen des Aufsatzbewertens und nicht nur die der Notenvorhersage erfüllt werden (Schneider & Zweig 2022, 287–289, vgl. auch Zweig 2023, 166ff.). Wir haben uns daher ganz grundlegend damit befasst, was alles zu einem gelingenden Bewertungsverfahren gehört, indem wir die sprechakttheoretischen Gelingensbedingungen des Aufsatzbewertens nach Austin offengelegt haben. Wichtig sind hier u. a. die Bedingungen B 1 und B 2, die darin bestehen, dass das Verfahren „korrekt“ und „vollständig“ durchgeführt werden muss (Austin 1975, 36–38). Zur korrekten und vollständigen Durchführung gehört im Fall der Aufsatzbewertung essentiell, dass neben der Notenvergabe auch eine dazu passende **Begründung** geliefert werden kann (Schneider & Zweig 2022, 287ff.). Dies ist aber beim E-Rater grundsätzlich nicht möglich: Er greift nämlich, semiotisch gesprochen, ausschließlich auf **Symptome** (vgl. Keller 2018, 155–168) zu, nicht aber auf **Qualitätskriterien**: Er wurde so programmiert, dass er u. a. die verwendeten Modalverben und bestimmte Arten von Nebensätzen pro Aufsatz zählt. Zudem findet in Bezug auf die jeweilige Essay-Aufgabe z. B. bei Toefl-Tests im Vorfeld ein maschinelles Training mit 250 bis 300 von Menschen bewerteten Aufsätzen statt (vgl. auch Burstein et al. 2013a: 61). Das System ‚lernt‘ dabei, welche Symptome essayartige Aufsätze aufweisen, die von Menschen mit einer guten Note bewertet wurden; ebenso solche, die schlechter bewertet wurden. Es ‚lernt‘, welche lexikalischen und syntaktischen Merkmale hier am stärksten mit einer bestimmten Note assoziiert sind. Dann erfolgt eine ‚eigenständige‘ maschinelle Vorhersage der Noten weiterer Essays zum selben Thema.

Der springende Punkt ist aber der folgende: Die Symptome, die der E-Rater ermittelt, könnten von Menschen nicht sinnvollerweise als Bewertungskriterien angegeben werden: Man könnte nicht überzeugend darlegen, dass ein Essay positiv bewertet wurde, **weil** er bestimmte Wörter und bestimmte syntaktische Konstruktionen enthält. Wollte man dies in Zukunft so tun, so würde man ein reines ‚teaching and learning to the test‘ (vgl. Anson & Perelman 2017, 282 f.) praktizieren. Nun mag es durchaus realistisch sein, dass dies in manchen Bildungskontexten tatsächlich schon so gehandhabt wird – man lernt dann nur ganz schematisch, ‚was der Lehrer hören will‘; man vermittelt umgekehrt nur Oberflächenphänomene ohne Tiefgang. Es

¹ Der vorliegende Aufsatz basiert auf zwei Vorträgen, die ich bei der GAL-Jahrestagung 2023 in Mainz sowie im Seminar von Katrin Lehnen in Gießen gehalten habe. Ich danke allen dort Beteiligten, insbesondere Katrin Lehnen und Helmuth Feilke, für wichtige Anregungen. Für weitere Denkanstöße sowie praktische Unterstützung danke ich Hans Lösener, Michael Rödel und Katharina Zweig.

ist bislang aber wohl keineswegs so, dass diese Praxis als erstrebenswert propagiert wird. Vielmehr werden Kriterien wie Kohärenz, argumentative Plausibilität, Wahrheitsgehalt, Originalität, Ästhetik aus guten Gründen weiterhin angeführt, und solche sachgemäßen Kriterien werden bezeichnenderweise auch im Patent des E-Raters genannt (Burstein et al. 2002). Jedoch greift der E-Rater bestenfalls indirekt darauf zu. Er analysiert „Kohäsion ohne Kohärenz“ (Schneider & Zweig 2022, 289), was dazu führt, dass Nonsense-Texte Bestnoten erhalten können (vgl. dazu Perelman 2020, zu den Hintergründen auch Burstein et al. 2013b). Die Notenvorhersage, die der E-Rater abliefern, liegt in vielen Fällen zwar recht nah bei der menschlichen Notengebung (vgl. etwa Rupp et al. 2019, Meyer et al. 2023), und in solchen Fällen korrelieren offenbar die vorhandenen Oberflächen-Symptome mit den erfüllten Qualitätskriterien. Dies ändert aber nichts daran, dass die Qualitätssymptome, die E-Rater in Anschlag bringt, als Bewertungsgründe nicht stichhaltig sind. Setzt man den E-Rater zur Bewertung ein, so ist eine Begründung nicht mehr gewährleistet. Die Begründung dient jedoch dazu, das Verfahren für die Zukunft zu stabilisieren; nur so können die kulturell verankerten Gelingensbedingungen des Bewertens aufrechterhalten werden.

Die Aufsatzbewertungs-KI kann allerdings, wie bereits angedeutet, unter günstigen Bedingungen für ‚normale‘ Essays die von ‚normalen‘ GutachterInnen zu erwartende Note recht genau vorhersagen. Sie kann also nicht bewerten, eine menschliche Bewertung auch nicht leistungsgleich substituieren, aber sie kann – wie wir in unseren beiden Aufsätzen herausgearbeitet haben – unter günstigsten Bedingungen eine Essaybewertung im Sinne eines Korrektivs unterstützen: Dazu muss die KI, was ja auch gängige Praxis ist, für jeden Essayprompt eigens mit 250–300 von Menschen benoteten Essays trainiert werden, wodurch Aufgaben, die nur in relativ kleinen Gruppen, z. B. in nur einer Schulklasse, bearbeitet werden, ohnehin ausscheiden. Zudem muss mindestens ein Mensch bei jedem Essay gutachten. Durch den Vergleich von vorhergesagter Note und Note des Menschen können dann gegebenenfalls *Ausreißer* identifiziert werden: Weicht bei solchen Ausreißern der Mensch vom Bewertungsschema ab? Oder wird die Symptom-Erfassung dem Essay vielleicht nicht gerecht? (vgl. Schneider & Zweig 2022, 290).

Ganz am Ende unseres deutschsprachigen Beitrags von 2022 haben wir folgenden Ausblick formuliert:

Für die Zukunft bleibt die Frage, ob der E-Rater so überzeugend werden könnte wie etwa DeepL, wenn er den Bewertungsprozess selbst erlernen würde, indem er massenhaft mit menschlichen Textbewertungen von Schüleraufsätzen versorgt wird. Dies würde allerdings voraussetzen, dass diese Textbewertungen während des maschinellen Lernprozesses des E-Raters mit den entsprechenden Aufsätzen verbunden werden. Dann könnte der E-Rater vielleicht eine textuelle Bewertung erzeugen, die genau auf einen konkreten neuen Aufsatz bezogen und abgestimmt wäre und sich dann vielleicht sogar – wie DeepL – einer „leistungsgleichen Substitution“ [...] annähern könnte. Aber ist ein solcher Prozess der Aufsatzbewertung überhaupt technisch möglich, wenn man bedenkt, dass die inhaltliche Beziehung zwischen Aufsatz und Bewertungstext viel komplexer und ‚loser‘ ist als die zwischen einem Text und seiner

Übersetzung? Und selbst wenn es möglich wäre, könnten wir auf menschliche Gutachter nicht verzichten, denn nur sie könnten auf Nachfrage die Begründung weiter erläutern und die Verantwortung für die Bewertung der Aufsätze übernehmen. (Schneider & Zweig 2022, 291f.)

Diese Überlegungen wurden formuliert, bevor der große Boom der LLMs, insbesondere der ChatGPT-Hype, begann, und wir ahnten damals noch nicht, wie verblüffend die von ChatGPT produzierten textlichen Ausgaben mitunter sein können. Es ist daher sehr naheliegend, vor dem Hintergrund unserer Befassung mit dem E-Rater die Frage zu stellen, welche Rolle ChatGPT in einem solchen Essay-Bewertungsverfahren spielen kann (vgl. auch Zweig 2023, 179–187). Kann ChatGPT unter Umständen die fehlende Notenbegründung leistungsgleich substituieren oder zumindest eine Hilfe dafür darstellen, eine solche Begründung zu formulieren?

2 Kann ChatGPT eine Essaybewertung begründen?

ChatGPT („Generative Pre-trained Transformer“) ist ein LLM-Chatbot, der mit riesigen Textkorpora trainiert wurde, auf der Basis statistischer Wahrscheinlichkeiten von Wortfolgen ‚eigene‘ Texte zu erzeugen. Er kann auf einen Prompt reagieren und dann einen Text generieren, der ähnlich ist wie diejenigen Texte, mit denen das neuronale Netz trainiert wurde. Ein bekanntes Problem ist dabei das sogenannte ‚Halluzinieren‘ bzw. die Unzuverlässigkeit im Hinblick auf Tatsachenbehauptungen. Hieran wurde in den letzten Monaten stark gearbeitet, und das Erzeugen völlig unwahrer Behauptungen ist deutlich geringer geworden, das Problem ist aber nach wie vor gegeben: Nach Austins Sprechakttheorie (1975, 92f.) ist hier insbesondere der „rhetische Akt“ betroffen, der darin besteht, dass auf etwas referiert wird (*reference*) und darüber etwas gesagt wird (*sense*). Oder, in der heute gängigen Terminologie der linguistischen Pragmatik: Es werden Referenz- und Prädikationsakte vollzogen.

Im Folgenden möchte ich in einer Art Werkstattbericht darlegen, welche Ergebnisse ich in Bezug auf Essayformulierung, -benotung und insbesondere Notenbegründung mit GPT3.5 und GPT4 erzielt habe, wobei das Hauptinteresse auf der Frage liegt, wie gut rhetische Akte hier substituiert werden können. Die Forschungsfragen dieser explorativen Studie lauten:

Kann ChatGPT eine menschliche Notenbegründung (näherungsweise) leistungsgleich substituieren? Größer gefragt: Reicht die massenhafte syntaktische Exemplifikation von Wörtern in realen Kontexten, mit der ChatGPT trainiert wird, unter Umständen (irgendwann) aus, um die mangelnde Referentialität auszugleichen?

2.1 Versuche mit GPT3.5 (August 2023)

Um diesen Fragen nachzugehen, habe ich GPT3.5 im August 2023 zunächst mit folgendem Prompt die Aufgabe gestellt, einen Aufsatz zu formulieren, wie es in Toefl-Tests üblich ist:

Prompt I:

„Überall, so scheint es, gibt es klare und positive Anzeichen dafür, dass die Menschen die Unterschiede zwischen den Menschen mehr respektieren.“ Wie zutreffend ist Ihrer

Meinung nach die oben geäußerte Ansicht? Schreiben Sie darüber einen Aufsatz von etwa 400-500 Wörtern. Verwenden Sie Gründe und/oder Beispiele aus Ihrer eigenen Erfahrung, Ihren Beobachtungen oder Ihrer Lektüre, um Ihren Standpunkt zu begründen.

Die Ausgabe von GPT3.5, die dann erfolgte, lässt sich wirklich sehen und ist im Anhang dieses Beitrags unter (A) abgedruckt. Es handelt sich um einen wohlstrukturierten Text, der auf die Essay-Aufgabe angemessen antwortet und mit Sicherheit als bestanden bewertet würde – wahrscheinlich im oberen Zweierbereich, wenn man deutsche Noten anwendet.

In einem zweiten Schritt habe ich GPT3.5 dann genau diesen Aufsatz zur Bewertung gegeben, und zwar mit folgendem Prompt:

Prompt 2:

Ein Bewerber hat bei einem Sprachtest folgenden Aufsatz geschrieben, um einen Studienplatz in einem deutschsprachigen Studiengang zu erhalten. Die Aufgabe bestand darin, folgende These zu bewerten: „Überall, so scheint es, gibt es klare und positive Anzeichen dafür, dass die Menschen die Unterschiede zwischen den Menschen mehr respektieren.“ Der Bewerber sollte dazu einen Aufsatz von ca. 400-500 Wörtern schreiben und erörtern, wie zutreffend seiner Meinung nach die oben geäußerte Ansicht ist. Er sollte dabei Gründe und/oder Beispiele aus seiner eigenen Erfahrung, seinen Beobachtungen oder seiner Lektüre verwenden, um seinen Standpunkt zu begründen. Der Aufsatz steht unten. Bitte bewerten Sie ihn mit einer Note zwischen 1 und 6, wobei Sie auch Noten wie 2 plus (1,7) oder 3 minus (3,3) vergeben können, und schreiben Sie eine Begründung für diese Rückmeldung. Sie sollte auch drei Tipps enthalten, wie sich der Bewerber verbessern kann. Das ist der Aufsatz: [...]

Den Aufsatz habe ich dann einfach in den Prompt 2 hineinkopiert und diesen Prompt im August 2023 mehrfach wiederholt. Es kamen damals immer Standardbegründungen heraus, die insofern an Zeitungshoroskope erinnerten, als sie auf fast jeden einigermaßen gelungenen Essay zutreffen würden. Plakativ gesprochen: Es handelte sich um Standardbewertungen, die in meinen damaligen Versuchen von ChatGPT immer mit *gut* (2) benotet wurden. Es erfolgten oberflächliche Formulierungen zu Kohärenz, Stil, sprachlicher Korrektheit usw.², ohne dass dabei erkennbar auf den eingegebenen Text zugegriffen wurde. Ein Beispiel einer solchen ChatGPT-Notenbegründung findet sich im Anhang unter (B). Interessant ist dort, dass die drei Verbesserungstipps, um die ich im Prompt gebeten hatte, hier doppelt vorkommen. Offenbar war der Chatbot ohnehin darauf trainiert, solche Tipps zu geben und diese wurden dann, wohl aufgrund der expliziten Aufgabenstellung, schematisch gedoppelt. Allerdings erfolgte nicht bei allen Versuchen eine solche Dopplung. Auffällig ist zudem, dass niemals dieselbe Ausgabe erfolgte, auch wenn ich denselben Prompt mehrfach hintereinander eingab. Dies ist für diejenigen, die mit der Funktionsweise von ChatGPT vertraut sind, nicht verwunderlich, denn der Chatbot ist immer mit einer sogenannten ‚Temperatur‘ eingestellt (vgl. unten S. 13); das

² Zu Bewertungskriterien von Aufsätzen aus deutschdidaktischer Sicht vgl. Rödel 2018, Kap. 4.3.

heißt, er sucht nicht die am wahrscheinlichsten folgenden Wortformen und Wortfolgen aus³, sondern solche, die zu den wahrscheinlichsten gehören. Würde man immer nur die allerwahrscheinlichsten folgen lassen, so entstünden zwar identische, aber völlig langweilige, unkreative Texte, die sozusagen mit minimalem Risiko erzeugt würden. Stellt man die Temperatur dagegen zu hoch ein, entstehen irgendwann völlig sinnlose Textgebilde.

In einem zweiten Schritt habe ich den von ChatGPT generierten Essay dann eigenhändig verschlechtert und dabei zwei Versionen erzeugt: Eine, in der *sprachlich-stilistische und logische* Fehler eingebaut wurden (siehe Anhang C, wo – nur zur Orientierung der LeserInnen dieses Beitrags – die sprachlich-stilistischen Fehler grün markiert sind, die logischen Fehler gelb), und eine zweite Version, die *sachlich-referentielle* Fehler enthält (siehe Anhang D, wo diese Fehler zur Orientierung gelb markiert sind). Mithilfe dieser beiden Versionen habe ich ChatGPT dann weiter getestet. Zunächst, indem ich wieder um Bewertungsbegründung und Benotung bat. Hierbei entstanden zunächst wieder ‚horoskopartige‘ Ausgaben, die inhaltlich kaum auf den Text zugegriffen und in der Benotung erratisch waren.

Deutlich interessanter wurde es, wenn ich ChatGPT dann nach einzelnen Bewertungsaspekten fragte, z.B. mit diesem Prompt, in den ich den sprachlich und logisch verschlechterten Essay hineinkopierte:

Prompt 3:

Welche sprachlichen und stilistischen Fehler enthält der folgende Text? [...]

Hier einige Beispiele aus einer ChatGPTs Antwort⁴:

- „unbestreitbar, das“ muss zu „unbestreitbar, dass“ korrigiert werden.
- „Respeckts“ sollte „Respekts“ sein.
- „sexuellem Orientierung“ sollte „sexueller Orientierung“ sein.
- „megagut“ ist umgangssprachlich und sollte ersetzt werden.
- „interkulturelles Geschwafel“ sollte wahrscheinlich „interkulturelles Verständnis“ sein.
- „Herausforderungen im Keim zu ersticken“ ist nicht ganz klar, es könnte besser ausgedrückt werden.
- „ein Klima der Intoleranz zu schaffen“ sollte „ein Klima der Toleranz zu fördern“ sein.

³ Dies ist natürlich wieder eine metaphorische, anthropozentrische Formulierung, denn der Chatbot sucht im buchstäblichen Sinne nichts aus, genauso wenig wie er etwas bewertet, voraussagt usw. Ob eine solche Ausdrucksweise **generell** problematisch ist, solange man sich über die Metaphorizität im Klaren ist, darüber lässt sich trefflich streiten. Manche Metaphoriken sind indes sprachphilosophisch so irreführend, dass ich sie im späteren Verlauf des vorliegenden Artikels infrage stelle – insbesondere betrifft dies die Begriffe ‚intelligent‘ und ‚Text‘: Chatbots produzieren keine **intelligenten Texte**, sondern **intelligible Texturen**. Was es damit auf sich hat, wird in Abschnitt 3 erklärt. Auf die Idee, hier den Begriff ‚intelligibel‘ zu verwenden, kam ich durch ein Gespräch mit dem Philosophen Ralf Becker, dem ich an dieser Stelle herzlich danke.

⁴ Es handelt sich hier und im Folgenden um wörtliche Zitate aus den ChatGPT-Ausgaben, die zur besseren Übersichtlichkeit farblich (blau) abgesetzt wurden. Ich habe alle Prompts (lila) mehrfach und an verschiedenen Tagen und Uhrzeiten des Augusts 2023 getestet. Die Qualität der Antworten war sehr unterschiedlich. Ich führe im Folgenden tendenziell eher die ‚besseren‘, jedenfalls die interessanteren Antworten an, um einen Einblick in das Leistungsvermögen von GPT3.5 zu diesem Zeitpunkt zu gewähren. Hieraus konkrete Schlussfolgerungen zu ziehen, ist mit Vorsicht zu genießen; gewisse Tendenzen lassen sich allerdings trotzdem herausarbeiten, solange man den Werkstattcharakter im Auge behält.

Wie die Beispiele zeigen, ist bereits GPT3.5 gut in der Lage, sprachliche Abweichungen (insbesondere orthografische und grammatische) zu identifizieren und die betreffenden Ausdrücke durch korrektere bzw. angemessenere zu ersetzen, wobei unklar ist, wie sich das technisch genau erklären lässt.⁵ Auch umgangssprachliche, informelle Ausdrücke werden als nicht-standardsprachlich markiert und Alternativen angeboten. Gleichzeitig fällt an den letzten beiden Beispielen auf, dass die Fehler (noch) nicht alle korrekt klassifiziert werden. Der idiomatische Ausdruck *im Keim zu ersticken* wird stilistisch moniert, obwohl er zwar metaphorisch, in einem Essay jedoch nicht unbedingt als unangemessen zu betrachten ist. Zu beanstanden ist an dieser Textstelle vielmehr die fehlende logische Stringenz, denn die im Ursprungstext dort stehende Infinitivphrase *zu fördern* entsprach der Kohärenz des Gesamttextes, während *im Keim zu ersticken* das Gegenteil ausdrückt. Analog dazu wird beim letzten Beispiel der Ausdruck *ein Klima der Intoleranz zu schaffen* zwar zurecht beanstandet und korrekt ersetzt, jedoch wird auch dieser Fehler falsch kategorisiert und nicht als logischer Widerspruch erkannt.

Nach logischen Fehlern in dieser Essay-Version wurde erst mit dem nächsten Prompt gefragt:

Prompt 4:

Welche logischen Fehler und Ungenauigkeiten enthält der folgende Text? [...]

Die Ausgabe von GPT3.5 enthielt diesmal u. a. folgende Feststellungen:

- Im Abschnitt „Gesellschaftliche Rückschritte“ wird die "sinkende Anerkennung der Bedeutung von Gleichberechtigung und Toleranz" als Fortschritt dargestellt, was inkorrekt ist. Es sollte wahrscheinlich „steigende Anerkennung“ heißen, um den Gedanken korrekt auszudrücken.
- Im Fazit wird die Formulierung „Klima der Intoleranz zu schaffen“ verwendet, was das Gegenteil dessen ist, was vermittelt werden soll. Es sollte heißen „Klima der Toleranz zu schaffen“.

Die Fehlerkategorisierung ist hier also korrekt, ebenso bei den weiteren Feststellungen, die ich hier nicht eigens zitiere.

Noch interessanter wird das Ganze, wenn nach der sachlich-referentiellen Korrektheit des Textes gefragt, also der rhetische Akt, fokussiert wird. Der Prompt hierzu lautete:

Prompt 5:

Nennen Sie mir alle Unwahrheiten, die in dem folgenden Text stecken: [...]

Die Antwort von GPT3.5 ist durchaus erstaunlich und wird hier vollständig wiedergegeben:

Es scheint, dass der Text absichtlich mit einigen Unwahrheiten und Fehlinformationen durchsetzt ist. Hier sind die identifizierten Unwahrheiten:

⁵ Die Maschine kann nicht wirklich ‚feststellen‘, dass ein Wort falsch geschrieben oder grammatisch falsch gebildet ist; dafür gibt es hier kein technisches Analogon (Zweig im persönlichen Gespräch, März 2024).

- **Verbot von nicht regierungsfreundlichen Zeitungen:** Im Text wird behauptet, dass Zeitungen, die nicht die Meinung der Bundesregierung vertreten, in Deutschland verboten seien. Das ist nicht wahr. In Deutschland gibt es Meinungs- und Pressefreiheit, und es gibt viele unabhängige Medien, die verschiedene Standpunkte vertreten.
- **Einfluss der Bundesregierung auf Medieninhalte:** Der Text behauptet, dass die Bundesregierung die Medien gesetzlich dazu verpflichtet hat, Vielfalt und Unterschiede darzustellen. Dies ist eine Übertreibung. Während es sicherlich Richtlinien und Gesetze gibt, die Diskriminierung verhindern sollen, ist die Darstellung von Vielfalt in den Medien nicht direkt von der Bundesregierung gesteuert.
- **Einmischung der Bundesregierung in die Schulen:** Der Text behauptet, dass die Bundesregierung in Deutschland für Schulen zuständig sei und nicht die Länder. Tatsächlich ist das deutsche Bildungssystem föderal aufgebaut, was bedeutet, dass die Bildungshoheit in erster Linie bei den Bundesländern liegt.
- **Kontakt zwischen Menschen aus der BRD und DDR:** Der Text erwähnt, dass der Kontakt zwischen Menschen aus der Bundesrepublik Deutschland (BRD) und der Deutschen Demokratischen Republik (DDR) immer stärker wird. Da die DDR 1990 aufhörte zu existieren, ist diese Aussage nicht korrekt.
- **Die Grünen als rechtsextrem eingestuft:** Die Aussage, dass die Grünen vom Verfassungsschutz als rechtsextrem eingestuft werden, ist definitiv falsch. Die Grünen sind eine politische Partei, die in Deutschland eine lange Geschichte hat und nicht als rechtsextrem eingestuft wird.
- **Zukünftiger Abriss der Berliner Mauer:** Der Text hofft darauf, dass die Berliner Mauer bald abgerissen wird. Die Berliner Mauer wurde jedoch bereits im Jahr 1989 weitgehend abgerissen.

Bitte beachten Sie, dass der Text möglicherweise satirisch oder humorvoll gemeint sein könnte und nicht unbedingt beabsichtigt, Fakten korrekt darzustellen.

Nicht nur werden hier einige meiner eingebauten inhaltlichen Fehler korrekt offengelegt, sondern es werden im Eingangs- und Satzsatz sogar Mutmaßungen über den Status und die Intention des Gesamttextes formuliert: Die „Unwahrheiten und Fehlinformationen“ könnten absichtlich eingebaut worden sein, der Text könnte „satirisch oder humorvoll gemeint“ sein. Andere Fehler jedoch wurden fast hartnäckig nicht erkannt: Auch bei genauer Nachfrage in den Prompts ‚behauptet‘ GPT3.5 unzutreffend, Angela Merkel habe im Bundestag für die ‚Ehe für alle‘ gestimmt, bei nochmaliger Nachfrage wird abgeschwächt, sie habe sich enthalten – was aber ebenfalls nicht zutrifft, denn die damalige Bundeskanzlerin hat 2017 dagegen gestimmt. Auffällig war bezüglich der ‚Rhetorik‘ von GPT zu diesem Zeitpunkt, dass sich der Chatbot bei kritischen Rückfragen mit einer – vermutlich im überwachten Feintuning gelernten – Phrase entschuldigt, seine Aussage korrigiert, aber weiterhin tendenziell Fehlinformationen produziert:

Es tut mir leid für die Verwirrung. Du hast recht, Angela Merkel hat sich nicht für die „Ehe für alle“ eingesetzt, sondern sich enthalten, als im Deutschen Bundestag über die Einführung

der „Ehe für alle“ abgestimmt wurde. Dies geschah im Jahr 2017. Ich entschuldige mich für die falsche Information in meiner vorherigen Antwort. Danke, dass du das klargestellt hast.

2.2 Versuche mit GPT4 (Januar 2024)

Im Januar habe ich die bereits verwendeten Prompts mit GPT4 erneut getestet. Während GPT3.5 im August 2023 mit Daten bis September 2021 trainiert war, speist sich GPT4 aus Daten bis April 2023. Zunächst wende ich wieder Prompt 2 (mit den GPT-Essay vom August 2023) an und bitte GPT4 um eine Gesamtbewertung. Im Gegensatz zu GPT3.5 wird nun erstmalig die im Prompt eröffnete Möglichkeit berücksichtigt, Zwischennoten anzuwenden: die Note lautet diesmal 1,7. Insgesamt lässt sich zudem festhalten, dass die Chatbot-Antwort immer noch einen gewissen Horoskop-Charakter aufweist, diesmal aber etwas stärkere inhaltliche Bezüge zu dem Essay, der zur Bewertung steht, erkennen lässt. Dies zeigt sich vor allem in dieser Formulierung:

Obwohl der Aufsatz eine ausgewogene Sicht bietet, könnte eine tiefere Analyse der komplexen Gründe für die anhaltende Diskriminierung und Vorurteile die Argumentation bereichern. Dies würde eine kritischere Auseinandersetzung mit dem Thema ermöglichen.

Die Referenz auf „die anhaltende Diskriminierung und Vorurteile“ liest sich als inhaltliche Verbindung zwischen der Bewertung und dem bewerteten Essay, was in dieser Weise bei GPT3.5 kein einziges Mal der Fall war. Jedoch bleibt die Gesamtbewertung insgesamt auch diesmal an der Oberfläche und ist voller Gemeinplätze, die auf sehr viele einigermaßen gelungene Aufsätze zutreffen.

Interessant ist vor allem der Vergleich bei den verschiedenen Fehler-Kategorien mit Bezug auf die Versionen des Textes, die von mir verschlechtert wurden – ich verwendete wieder die Versionen, die im Anhang unter Punkt (C) und (D) zu finden sind. Bezüglich der sprachlich-stilistischen Fehler (Prompt 3) ergaben sich sehr ähnliche Antworten wie bei GPT3.5: Diese werden gut erkannt. Die Kategorisierung der Fehler war noch etwas genauer als bei GPT3.5, jedoch auch nicht immer korrekt.

Hinsichtlich der logischen Fehler und Ungenauigkeiten war eine Antwort von GPT4 so erstaunlich präzise und klar auf den Essay bezogen, dass ich sie hier vollständig zitiere. In fünf gut gegliederten Punkten werden logische Schwächen des Textes glasklar analysiert:

1. Widersprüchliche Argumente:

Der Text präsentiert an verschiedenen Stellen widersprüchliche Aussagen. Zum Beispiel erwähnt der Autor, dass rechtliche Maßnahmen gegen Diskriminierung eine „sinkende Anerkennung der Bedeutung von Gleichberechtigung und Toleranz“ widerspiegeln, was logisch unvereinbar erscheint, da solche Maßnahmen normalerweise als Fortschritt in Richtung Gleichberechtigung und Toleranz angesehen werden.

2. Fehlinterpretationen und Missverständnisse:

Der Autor behauptet, dass Medien dazu beitragen, „Stereotypen zu zementieren und Empathie zu ersticken“, was eine negative Rolle der Medien impliziert. Diese Aussage steht

im Widerspruch zu der üblichen Auffassung, dass Medien eine positive Rolle bei der Förderung von Diversität und Verständnis für unterschiedliche Lebenserfahrungen spielen können.

3. Ungenauigkeiten in der Darstellung:

Der Abschnitt über Bildung spricht von der Integration von „interkulturellem Geschwafel“ in Lehrpläne. Diese Formulierung ist abwertend und ungenau, was die Seriosität des Arguments mindert.

4. Negativ konnotierte Sprache:

Der Gebrauch negativ konnotierter Begriffe wie „Geschwafel“ oder die Formulierung „lernen zu früh“ suggeriert eine negative Bewertung der Bildungsinitiativen, was der üblichen positiven Bewertung solcher Bildungsmaßnahmen widerspricht.

5. Inkonsistenz in der Schlussfolgerung:

Das Fazit, dass gesellschaftliche Fortschritte und Bildungsbemühungen „zweifelloso dazu beitragen, ein Klima der Intoleranz zu schaffen“, steht im Gegensatz zu den üblichen Annahmen über den Einfluss dieser Faktoren auf die Förderung von Toleranz und Verständnis.

Diese Antwort zeigt das Potenzial, das GPT im Hinblick auf eine Bewertungsbeurteilung hat – es sei jedoch nochmals (vgl. Fußnote 4) darauf hingewiesen, dass ich die Prompts an verschiedenen Tagen und zu verschiedenen Uhrzeiten wiederholt habe und die Qualität der Antworten auch bei GPT4 erstaunlich divergierte: Von völlig oberflächlichen, auch fehlerhaften, bis hin zu präzisen und brauchbaren Antworten wie der gerade zitierten, die sich am obersten Rand der Ergebnisse bewegt.

Nachdem die Gesamtbewertung des ‚Standardaufsatzes‘ bei GPT4 geringfügig aussagekräftiger war als bei GPT3.5, habe ich dann noch getestet, wie GPT4 den von mir mit Unwahrheiten gespickten Text bewertet und benotet. Ich habe dem Chatbot also Version (D) mit Prompt 2 zur Bewertung gegeben, ohne dabei zunächst auf einzelne Fehlerkategorien zu verweisen. Hier eine, wirklich erstaunliche, Antwort des Chatbots:

Bewertung des Aufsatzes: Note 5 (mangelhaft)

Begründung:

1. Faktische Ungenauigkeiten und historische Fehler: Der Aufsatz enthält mehrere gravierende Fehler. Angela Merkel war Bundeskanzlerin, nicht Bundespräsidentin, und Willy Brandt war in den 1970er Jahren Bundeskanzler und nicht in der Zeit, als die "Ehe für alle" eingeführt wurde. Zudem ist die Behauptung, dass in Deutschland Zeitungen, die nicht die Meinung der Bundesregierung vertreten, verboten sind, sachlich falsch und irreführend. Diese Fehler zeigen ein mangelndes Verständnis für das politische System Deutschlands und seine Geschichte.

2. Thematische Abweichungen und unzutreffende Darstellungen: Der Bewerber erwähnt die Grünen als rechtsextreme Gruppe, was historisch und politisch inkorrekt ist. Ebenso ist die Erwähnung der „Berliner Mauer“ und der „Reichsbürger“ nicht nur unzutreffend, sondern auch irrelevant für das Thema des Aufsatzes. [...] [Es folgen noch „Sprachliche Mängel“ & „Tipps zur Verbesserung“]

Erstaunlich fand ich insbesondere die Feinunterscheidung zwischen den beiden fettgedruckten Kategorien und fragte daher diesbezüglich in einem weiteren Prompt nach:

Prompt 6:

Warum haben Sie bei der Begründung die Punkte 1. und 2. unterschieden? Worin unterscheiden sich die beiden Kategorien? Bezieht sich nicht beides auf historisch falsche Darstellungen?

Die Antwort von GPT4 lautet:

Sie haben Recht, dass sowohl Punkt 1 als auch Punkt 2 meiner Begründung sich auf historisch falsche Darstellungen beziehen. Der Unterschied zwischen den beiden Punkten liegt jedoch in der Art und Weise, wie diese Fehler sich auf den Aufsatz auswirken:

1. Faktische Ungenauigkeiten und historische Fehler: Hier geht es um spezifische Fehler in Bezug auf historische Fakten und Personen, wie die falsche Benennung von Angela Merkel als Bundespräsidentin oder die fehlerhafte Einordnung von Willy Brandt in einen falschen zeitlichen Kontext. Diese Fehler zeigen ein grundlegendes Missverständnis der deutschen politischen Geschichte und beeinträchtigen die Glaubwürdigkeit der Argumentation, da sie auf ungenauen oder falschen Informationen basieren.

2. Thematische Abweichungen und unzutreffende Darstellungen: In diesem Punkt geht es um inhaltliche Fehler, die über reine Faktenfehler hinausgehen und thematisch irrelevante oder inkorrekte Aussagen beinhalten, wie die Behauptung, dass in Deutschland Zeitungen verboten sind, die nicht die Meinung der Bundesregierung vertreten, oder die falsche Darstellung der Grünen als rechtsextreme Gruppe. Diese Fehler zeigen nicht nur eine mangelnde Kenntnis der Fakten, sondern auch eine unangemessene und irreführende Darstellung von politischen und gesellschaftlichen Sachverhalten, die für das Verständnis des Themas wesentlich sind.

Insgesamt beziehen sich beide Punkte zwar auf historische und faktische Ungenauigkeiten, aber sie betreffen unterschiedliche Aspekte der Argumentation und Analyse im Aufsatz. Der erste Punkt bezieht sich auf konkrete Faktenfehler, während der zweite Punkt eine inhaltliche und thematische Verfehlung sowie eine verzerrte Darstellung der Realität umfasst.

Diese Antwort ist zweifellos als intelligente Erwiderung lesbar. Im Gegensatz zu GPT3.5, der auf Nachfragen von mir immer mit ‚Entschuldigungen‘ und sofortigem inhaltlichen ‚Einknicken‘ reagierte, erfolgt hier eine plausible Erläuterung der kategorialen Differenzierung, die das System vorgenommen hat. Die Antwort ist inhaltlich sehr klar auf den zu bewertenden Ausgangstext und auch auf die dann erfolgte GPT-Bewertung bezogen.

Nach der zitierten, plausiblen GPT4-Antwort bezüglich der beiden differenzierten Kategorien fragte ich mit zwei weiteren Prompts noch weiter:

Prompt 7:

Welche sachlichen Fehler sind in dem Text noch enthalten? Einige hatten Sie schon genannt.

Neben den bereits genannten führt GPT4 nun noch folgende Punkte an (Zitate aus der Antwort):

- Willy Brandt war Bundeskanzler in den 1970er Jahren und konnte daher Angela Merkel nicht bei der Einführung der „Ehe für alle“ unterstützen.
- Die „Ehe für alle“ wurde in Deutschland im Jahr 2017 legalisiert, nicht 2015.
- Die Hoffnung auf den Abriss der Berliner Mauer ist anachronistisch, da die Mauer bereits 1989 gefallen ist. Diese Aussage zeigt ein Missverständnis der jüngeren deutschen Geschichte.

Diese Kritikpunkte sind natürlich alle zutreffend und recht präzise formuliert. Was wiederum – wie schon bei GPT3.5 – fehlt, ist eine Bezugnahme darauf, dass Angela Merkel die ‚Ehe für alle‘ gar nicht vorangetrieben hat, im Bundestag sogar dagegen gestimmt hat. Nach einer nochmaligen Rückfrage zu diesem Punkt liefert GPT4 jedoch eine ausführliche und historisch einwandfrei Darstellung dieses Sachverhalts – von ‚Halluzinieren‘ keine Spur mehr.

Hinzuzufügen ist allerdings, dass die GPT4-Antworten auch in Bezug auf den verschlechterten Essay (D) insgesamt wenig konstant waren: Bei Wiederholungen des Prompts 2 mit diesem Essay gab es zum Teil wieder viel oberflächlichere Antworten und sehr verschiedene Noten, die von 3 bis 5 reichten.

2.3 Zwischenfazit zu den Antworten von ChatGPT3.5 und 4 im Vergleich

Vergleichend lässt sich nun Folgendes festhalten. Die Antworten, die bereits GPT3.5, erst recht GPT4 auf meine Prompts gegeben haben, sind zum Teil verblüffend, zum Teil sicherlich auch hilfreich, wenn man zu einer Notenbegründung gelangen möchte. Dies betrifft weniger die Gesamtbewertung, die GPT abgeliefert: Diese ist bei GPT3.5 durchgängig oberflächlich und zu wenig spezifisch; bei GPT4 greift sie anscheinend etwas mehr auf den zu bewertenden Text zu, ist aber ebenfalls weit davon entfernt, eine versierte menschliche Bewertung leistungsgleich zu ersetzen. Dies schließt in keiner Weise aus, dass viele menschliche Benotungsbegründungen auch nicht mehr in die Tiefe gehen als GPT-Begründungen – der Unterschied besteht aber darin, dass Menschen, wenn sie kompetent sind und sich Mühe geben, deutlich verlässlichere und aussagekräftigere Begründungen formulieren **können**. GPT3.5 bewertete Standardtexte mit der Standardnote gut (2). Bei Texten, die sehr mit Fehlern gespickt waren, schwankte er in der Notengebung. Wenn ich ihn auf Fehler aufmerksam machte, änderte er sofort seine Benotung und auch die dazugehörige Begründung. Salopp gesprochen: Er knickte gegenüber Kritik sofort ein. GPT4 war bei der Notengebung in Bezug auf ein und denselben Text schon beim Standardtext schwankend, bei den fehlerhaften Versionen erst recht. Er gibt häufig andere Noten für ein und denselben Text, die Begründungen dafür sind im Hinblick auf den Tiefgang sehr unterschiedlich: Manche lesen sich so, als ob sie tatsächlich auf den zu bewertenden Essay zugegriffen. Am erstaunlichsten war dies bei der oben wiedergegebenen und diskutierten Begründung, die zwei ad hoc gebildete Kategorien enthielt, welche GPT dann auch nach meiner Rückfrage plausibel erläuterte.

Die ‚Fähigkeit‘, sprachliche und stilistische Normabweichungen zu erkennen, wenn im Prompt konkret danach gefragt wird, war bei GPT3.5 bereits gut entwickelt. Bei der Frage nach logischen Fehlern war dies ebenfalls der Fall, wobei die Ergebnisse von GPT4 hier insgesamt etwas überzeugender waren und auch weniger Fehler in Bezug auf die Fehlerkategorisierung erfolgten. Am größten war der Unterschied bei der Frage nach den unwahren Aussagen in den Texten. Hier halluziniert GPT4 Anfang 2024 deutlich weniger als GPT3.5 im August 2023, wobei hinzuzufügen ist, dass auch GPT3.5 zwischenzeitlich in dieser Hinsicht verbessert wurde.

Insgesamt ist durchaus ein Fortschritt erkennbar, wenn man GPT3.5 (August 2023) mit GPT4 (Januar 2024) vergleicht. Insgesamt bleiben die Antworten in der Qualität jedoch schwankend und damit unzuverlässig, wobei die besten Ausgaben von GPT4 sich durchaus wie gelungene menschliche Gutachten lesen. Dies war bei GPT3.5 – zumindest in meinem Setting – noch nicht der Fall.

Die starken Schwankungen hängen – wie eingangs bereits angedeutet – damit zusammen, dass bei ChatGPT immer eine bestimmte sogenannte ‚Temperatur‘ eingestellt wird. Die Temperatur gibt vor, wie oft das Modell nicht immer das wahrscheinlichste, sondern auch einmal das zweit- oder drittwahrscheinlichste Wort als Vorhersagen nutzen soll (vgl. Wolfram 2023, 2). Wird hier eine sehr niedrige Temperatur, z.B. 0,1, gewählt, sind die Antworten zwar recht sicher, da auf extrem hoher Wahrscheinlichkeit beruhend, gleichzeitig aber relativ un kreativ und langweilig. Auch die Schwankungen in den Antworten sind beim selben Prompt zwar geringer⁶, wenn eine niedrige Temperatur gewählt wird, der Unterschied zu menschlichen Texten ist dann aber besonders deutlich. Insofern kann die Lösung auch nicht darin bestehen, immer die niedrigstmögliche Temperatur einzustellen. Wird allerdings eine sehr hohe Temperatur (bis zu 2,0) gewählt, dann werden die Texte irgendwann so ‚kreativ‘, dass sie nicht mehr als sinnvoll zu verstehen sind. Nach Wolfram hat sich eine Temperatur von 0,8 für die Essay-Generierung in der Praxis als optimal erwiesen (vgl. Wolfram 2023, 2).⁷

Trotz der Schwankungen lässt sich festhalten, dass bei GPT4 viel weniger Faktenfehler auftreten, was nicht zuletzt damit zusammenhängt, dass diese Version mit aktualisierten Trainingsdaten (bis April 2023) gefüttert ist, während die Trainingsdaten von GPT3.5 nur bis September 2021 gehen. Dennoch ist GPT4 ebenfalls nicht darauf ausgelegt, belastbare Auskünfte zu Faktenfragen zu geben. Auch wenn diese Version deutlich weniger halluziniert als GPT3.5, so war auch auf ihre historischen Behauptungen, selbst bei sehr bekannten Sachverhalten, keineswegs Verlass. Schon bei denselben Anfragen am selben Tag erfolgten zum Teil ganz andere, zum Teil fehlerhafte Antworten: z. B., dass Angela Merkel zu dem Zeitpunkt nicht

⁶ Ich habe dies mit Prompt 2 und der von mir festgelegten Temperatur 0,1 getestet. Auch hier waren die Antworten auf denselben Prompt nicht gleich.

⁷ Der Grad an Kreativität, der wünschenswert ist, hängt natürlich stark von der Textsorte und ihren Ansprüchen ab. 0,8 ist für Sachtexte bereits ein recht hoher Wert, der eine schon recht ausgeprägte Kreativität erwarten lässt. Als mittlere Werte werden häufig 0,5 bis 0,7 angenommen. Die Standardeinstellung von GPT soll 0,7 sein. Vgl. <https://www.german.ai/wissen/temperatur> (letzter Zugriff: 9. März 2024).

mehr Bundespräsidentin und dass Willy Brandt in den 1980ern **noch** Bundeskanzler gewesen sei.

Kirsten Schindler beschreibt die Kontingenz der Ausgabebedingungen von ChatGPT folgendermaßen:

*Ob die Ausgabe richtig ist, hängt vom Zufall, der Repräsentation des Inhalts im Material der Trainingsdaten, der Größe der Trainingsdaten und [den] Möglichkeiten zur Generalisierbarkeit ab. In den Fällen, in denen es nur wenige Trainingsdaten gibt oder Informationen vollständig unbekannt sind, sind die Ausgaben eher schlecht bzw. falsch. Für den Output ist also in besonderer Weise die Qualität des Inputs entscheidend, das gilt sowohl für die Trainingsdaten als auch für die von den Nutzer*innen formulierten Anfragen. (Schindler 2023, 5)*

Da sich die Ungenauigkeiten, wie gesagt, teilweise sogar bei sehr bekannten Fakten zeigen, ließe sich hier noch hinzufügen, dass auch die eingestellte ‚Temperatur‘ und der Zeitpunkt der Anfrage eine große Rolle spielen, wobei Schindler vor allem letzteren Aspekt mit ihrem Verweis auf „Zufall“ durchaus abdeckt.

3 Schlussdiskussion: Welche Rolle kann ChatGPT in der Aufsatzbewertung spielen?

Nehmen wir nun noch einmal eine etwas grundlegendere Haltung ein und betrachten das Gesagte aus einer größeren Flughöhe. Wie aus vielen Statements, wissenschaftlichen Untersuchungen und Systembeschreibungen zu ChatGPT bekannt ist, kann der Chatbot mit immer größerer Präzision die Wahrscheinlichkeit vorhersagen, welche Wörter, Phrasen, Sätze auf welche folgen.⁸ Die textuellen Ergebnisse, die dabei entstehen, sind zum Teil so verblüffend, dass auch Informatiker, die das System technisch maximal durchdringen, in letzter Instanz nicht erklären können, wie der konkrete Output zu erklären ist. Wie hilfreich ein solches LLM ist, muss im Hinblick auf verschiedene Domänen und ihre jeweiligen Anforderungen spezifisch beantwortet werden. Beim Thema ‚Aufsatzbewertung und Notenbegründung‘ lässt sich festhalten: Als **Werkzeug** kann ChatGPT dabei unterstützen, Essays zu Standardthemen zu bewerten; insbesondere kann es bei spezifisch gesetzten Prompts helfen, verschiedene Arten von Fehlern und sonstigen Schwächen in Texten zu finden, der Chatbot kann dann zu einer Art digitalem „Sparringspartner“ (vgl. Müller & Fürstenberg 2023, 341) für die Bewertenden werden. Auch positive Aspekte kann es manchmal verblüffend gut herausarbeiten. Es ist in seinen Antworten jedoch (noch) **zu instabil**, um eine **kriteriengeleitete Gesamtschau** zu liefern. Wie bereits im vorigen Abschnitt angesprochen, ließe sich natürlich diskutieren, wie sich die Ausgaben, die das System liefert, qualitativ im Vergleich zu realen menschlichen

⁸ Da es mir im vorliegenden Beitrag vor allem um eine linguistische und sprachphilosophische Analyse des **Outputs** von ChatGPT ging, habe ich auf eine technische Beschreibung weitgehend verzichtet. Eine verständlich formulierte und gleichwohl detaillierte Beschreibung aus informatischer Sicht ist z.B. Wolfram 2023. Aus linguistisch-kulturwissenschaftlicher Perspektive hilfreich ist Müller & Fürstenberg (2023, 328-335), wo die Funktionsweise von ChatGPT konzis dargestellt wird; eine verständliche und klare Kurzeinführung in die technischen Grundlagen bietet auch Schindler (2023, 3-7).

Gutachten verhalten. Gerade bei sehr standardisierten Essay-Aufgaben sind die Outputs von Mensch und Maschine hier wahrscheinlich recht ähnlich.

Ich möchte daher zum Abschluss die grundlegendere Frage wieder aufgreifen, die bereits am Anfang von Abschnitt 2 aufgeworfen und zunächst in Kurzform formuliert wurde: Kann ChatGPT eine menschliche Notenbegründung für einen Essay leistungsgleich substituieren? Isoliert man den phatischen Akt aus dem Austin'schen Sprechaktmodell, so lässt sich sagen, dass der Chatbot einen phatischen Akt bis zu einem gewissen Grad durchaus ersetzen kann; er kann Wörter und Sätze den grammatischen Regeln einer Sprache gemäß produzieren. Der zweite Teil der Austin'schen Definition des phatischen Aktes besagt aber darüber hinaus, dass Menschen solche Wörter und Sätze nicht nur produzieren, sondern sie auch **als** Wörter und Sätze der betreffenden Sprache produzieren können (vgl. Austin 1975, 92, vgl. hierzu auch Schneider 2008, 166f.). Dies kann nur dann vollumfänglich erfolgen, wenn der phatische Akt in einen rhetischen Akt eingebunden ist. Man verwendet Wörter, indem man sie auf **Kontexte** bezieht, indem man über etwas redet und darüber etwas sagt (oder schreibt). Dies kann der Chatbot aus ganz prinzipiellen Gründen nicht. Er nimmt nicht handelnd an Sprachspielen und damit auch nicht an der sozial geteilten Praxis des Regelfolgens im Sinne des späten Wittgenstein teil. Auf den Punkt gebracht: ChatGPT beherrscht Kontexte, aber keine Kontexte im handlungstheoretischen Sinne. Er fügt nicht über **Urteilkraft**, denn diese setzt voraus, dass man Erfahrungen in der Welt macht, dass man lernt, im Einzelfall zu entscheiden, ob etwas unter einen bestimmten Begriff fällt oder nicht. Dies ist nur möglich, wenn man eine leiblich-interaktionale Sozialisation erlebt, in der man, wie Wilhelm von Humboldt es so treffend beschrieben hat, die „Verstehbarkeit“ seiner eigenen Worte an anderen Menschen „versuchend“ prüft (Humboldt, Grundzüge, 377). Dies ist die Voraussetzung dafür, von Intelligenz in einem Sinne zu sprechen, welcher der hermeneutischen Tradition standhält. Die Leistungen, die der Chatbot erbringt, sind dadurch keineswegs geschmälert, nur muss der kategoriale Unterschied gesehen werden: ChatGPT erzeugt – so möchte ich dies auf den Begriff bringen – **intelligible Texturen**. Die entstehenden Produkte beruhen nicht direkt auf Sprechhandlungen mitsamt der dafür nötigen Urteilkraft, sondern auf statistischer Berechnung riesiger Datenmengen. Um dies zu verdeutlichen, gebrauche ich hier den von Christian Stetter (1997, 295ff.) eingeführten Begriff ‚Textur‘: Texturen sind materielle Zeichengebilde, also z. B. Buchstaben auf Papier, die erst dadurch, dass sie gelesen und verstanden werden, zu Texten werden. Insofern lässt sich sagen, dass ChatGPT intelligible Texturen erzeugt – **Gebilde, die sich als intelligente Texte lesen und interpretieren lassen**. Da das System mit (mehr oder weniger) intelligenten Texten von Menschen trainiert wurde, sind seine Ausgaben zwar ebenfalls potentiell intelligibel, jedoch insofern nicht intelligent, als in ihnen auf statistischer Basis Wortfolgen erzeugt, aber keine rhetischen Akte vollzogen werden, bei denen im Einzelfall schon vom Sprecher/Schreiber darüber entschieden wird, ob ein bestimmtes Prädikat auf ein bestimmtes Referenzobjekt zutrifft oder nicht. Dies von Fall zu Fall im freien Umgang mit gelernten Regeln entscheiden zu können, bildet nach Kant (KrV B, 172ff.) den Kern menschlicher Urteilkraft und damit auch einen entscheidenden Faktor dessen, was wir Intelligenz nennen.

In diesem Sinne haben Bermes und Dörpinghaus recht, wenn sie in der FAZ schreiben:

Es ist grundfalsch, das Leistungsvermögen von ChatGPT kleinzureden. Wer hier kleinlich nach ‚Fehlern‘ in den Antworten sucht, bleibt sprachlos, wenn sie behoben sind. (Bermes/ Dörpinghaus, FAZ Nr. 91, 19.04.2023, S. 4)

Sprachlos müssen wir dann nicht bleiben, wenn wir den kategorialen Unterschied zwischen intelligiblen Texturen und intelligenten Texten anerkennen, ohne damit den Wert der intelligiblen Texturen in Abrede zu stellen. Vor diesem Hintergrund können wir auch unsere komplexer formulierte empirische Frage aus Abschnitt 2 wiederholen: Reicht die massenhafte syntaktische Exemplifikation (vgl. Schneider 2015, 134f.) von Wörtern in realen Kontexten, mit der ChatGPT trainiert wird, unter Umständen irgendwann aus, um die mangelnde Referentialität, d. h. die mangelnden Kontexte, auszugleichen – ohne dass auffällt, dass keine echte Referenz vorliegt? Dies wäre über einen längeren Zeitraum in Bezug auf verschiedene Textsorten (Standardessay, Sachbuchtext etc.) und in unserem Fall speziell in Bezug auf die Bewertung von Exemplaren dieser Textsorten systematisch zu prüfen. Hierin steckt meines Erachtens ein reiches Forschungsprogramm, welches das Leistungsvermögen von LLMs auslotet, ohne die qualitativen Unterschiede zum Menschen aus Blick zu verlieren.

Vor allem: Mit der Frage nach der Urteilskraft ist immer wieder auch die Frage nach **Verantwortung** verbunden. Wer steht für die Wahrheit, Richtigkeit und Angemessenheit einer Aussage, eines Textes gerade? Auch die Frage der **Autorschaft** wird in diesem Zusammenhang in Bezug auf die Nutzung von KI wieder sehr virulent. Ganz neu ist das hier gemeinte Problem natürlich nicht: Praktiken des Copy and Paste spielen in einer Kultur der Digitalität (vgl. Stalder 2016) schon lange eine große Rolle, ebenso die Nutzung von Rechtschreib- und Synonymen-Programmen. Auch das Ghostwriting ist im Bildungsbereich, gerade an den Hochschulen, ein wichtiger Faktor. Für meine Argumentation hier entscheidend ist die Frage, wer den Text letztlich verantwortet. Und dies ist bei allen genannten Fällen klar: Es handelt sich jeweils um die Person (bzw. die Personengruppe), die den Text als den ihrigen einreicht. Der Ghostwriter lässt sich wohl kaum offiziell verantwortlich machen, wenn die Note nicht so ist wie erwartet und gewünscht, denn: Einen Ghostwriter einzusetzen verstößt bekanntermaßen gegen die Regeln des Sprachspiels ‚schriftliche Abschlussarbeit‘; das Risiko ist daher ganz von der Person zu tragen, die ihn beauftragt hat.

Bei ChatGPT kommt nun noch die Möglichkeit unbeabsichtigter Plagiate ins Spiel. Da der Chatbot (noch) nicht vertretbar mit Quellen umgehen kann und diese ja auch nicht offenliegen, muss jede Autorin und jeder Autor damit rechnen, dass Plagiate produziert werden, für die er oder sie dann geradestehen muss, auch wenn es unbeabsichtigt und unbemerkt geschieht. Ein Vergleich mit KI-Übersetzungsprogrammen wie DeepL ist hier klärend: Bei DeepL-Übersetzungen besteht die Frage normalerweise nur darin, ob der Ursprungstext gut und angemessen übersetzt wurde, nicht aber, ob der Ursprungstext der Wahrheit entspricht, Fakten korrekt wiedergibt usw. Letztere Aspekte beziehen sich auf den Ursprungstext und dessen Urheber. Bei ChatGPT dagegen ist die Nutzung der zugrundeliegenden Texte, derer

sich das System bedient, völlig intransparent und die Frage der Verantwortung ist daher mit deutlich größeren Risiken verbunden.

Gemeinsam ist Chatbot und Übersetzungsprogramm, dass sie dauerhaft auf menschlichen Input angewiesen bleiben. Beide müssen mit menschlichen Texten trainiert werden. Eine britische Studie von Shumailov et al. (2023) zeigt eindrucksvoll, dass ChatGPT kollabiert, wenn man das System in wiederholten Schleifen immer wieder mit maschinell generiertem Input füttert. Ab der ca. zehnten Stufe produziert der Chatbot keine intelligiblen Texturen mehr, sondern nur noch Unsinn. Auf menschliche Urteilskraft können wir nicht verzichten.

LITERATUR

Anson, Chris M. & Les Perelman. 2017. Machines can evaluate writing well. In Cheryl E. Ball & Drew M. Loewe (Hgg.): *Bad Ideas About Writing*, 278–286. Morgantown/WV: West Virginia University Libraries. Digital Publishing Institute (open access textbook). <https://textbooks.lib.wvu.edu/badideas/badideasaboutwriting-book.pdf> (Letzter Zugriff: 18.03.2024).

Austin, John L. 1975. *How to Do Things with Words*. 2. Aufl. Oxford: Oxford University Press.

Bermes, Christian & Andreas Dörpinghaus. 2023. Wer hat Angst vor ChatGPT? In: FAZ, Nr. 91, 19. April 2023, 4.

Burstein, Jill C., Lisa Braden-Harder, Martin S. Chodorow, Bruce A. Kaplan, Karen Kukich, Chi Lu, Donald A. Rock & Susanne Wolff. 2002. US 6,366,759 B1 [United States Patent, April 2, 2002: System and method for computer-based automatic essay scoring].

Burstein, Jill C., Joel Tetreault & Nitrin Madnani. 2013a. The E-rater Automated Essay Scoring System. In Mark D. Shermis & Jill Burstein (Hgg.), *Handbook of Automated Essay Evaluation. Current Applications and New Directions*, 55–67. London: Routledge

Burstein, Jill, Joel Tetreault, Martin Chodorow, Daniel Blanchard & Slava Andreyev. 2013b. Automated Evaluation of Discourse Coherence Quality in Essay Writing. In Mark D. Shermis, Jill Burstein (Hgg.), *Handbook of Automated Essay Evaluation. Current Applications and New Directions*, 267–280. London: Routledge.

Humboldt, Wilhelm von. [1968]. Grundzüge des allgemeinen Sprachtypus. In: Ders.: *Gesammelte Schriften*. Hrsg. von A. Leitzmann u.a. Berlin 1903ff. Photomechanischer Nachdruck Berlin 1968, Band 5. Zitiert als *Grundzüge*.

Janich, Peter. 2006. *Was ist Information? Kritik einer Legende*. Frankfurt am Main: Suhrkamp.

Janich, Peter. 2015. *Handwerk und Mundwerk. Über das Herstellen von Wissen*. München: Beck.

Kant, Immanuel. [1968]. Kritik der reinen Vernunft. In Ders.: Werke in 10 Bänden, hrsg. von Wilhelm Weischedel. Darmstadt: Wissenschaftliche Buchgesellschaft. Zitiert als KrV B.

Keller, Rudi 2018. Zeichentheorie. Eine pragmatische Theorie semiotischen Wissens. 2., durchgesehene Auflage. Tübingen: Narr.

Meyer, Jennifer, Thorben Jansen, Johanna Fleckenstein, Stefan Keller & Olaf Köller. 2023. Machine Learning im Bildungskontext: Evidenz für die Genauigkeit der automatisierten Beurteilung von Essays im Fach Englisch. Zeitschrift für Pädagogische Psychologie 37(1), 203-2014. <https://doi.org/10.1024/1010-0652/a000296>

Müller, Hans-Georg & Maurice Fürstenberg. 2023. Der Sprachgebrauchsautomat. Die Funktionsweise von GPT und ihre Folgen für Germanistik und Deutschdidaktik. Mitteilungen des Deutschen Germanistenverbandes 70(4), 327-345.

Perelman, Les. 2020. The BABEL Generator and E-Rater: 21st Century Writing Constructs and Automated Essay Scoring (AES). The Journal of Writing Assessment 13(1). <http://journalofwritingassessment.org/article.php?article=145> (Letzter Zugriff: 18.03.2024).

Rödel, Michael. 2018. Interpretationsaufsätze schreiben: Ein Handbuch. Baltmannsweiler: Schneider Hohengehren.

Rupp, André A., Jodi M. Casabianca, Maleika Krüger, Stefan Keller & Olaf Köller. 2019. Automated Essay Scoring at Scale: A Case Study in Switzerland and Germany. TOEFL Research Report Series and ETS Research Report Series ISSN 2330-8516, No. RR-19-12.

Schindler, Kirsten. 2023. ChatGPT oder Überlegungen zu den Veränderungen des Schreibens in der Schule. MiDU 5(2), 1-22. <https://doi.org/10.18716/OJS/MIDU/2023.2.5>

Schneider, Jan Georg. 2008. Spielräume der Medialität. Linguistische Gegenstandskonstitution aus medientheoretischer und pragmatischer Perspektive (= Linguistik – Impulse & Tendenzen 29). Berlin & New York: de Gruyter.

Schneider, Jan Georg. 2015. Syntaktische Schemabildung – zeichentheoretisch betrachtet. In Christa Dürscheid & Jan Georg Schneider (Hrsg.), Handbuch Satz, Äußerung, Schema (= HSW 4), 125–151. Berlin & Boston: de Gruyter.

Schneider, Jan Georg & Katharina A. Zweig. 2022. Ohne Sinn. Zu Anspruch und Wirklichkeit automatisierter Aufsatzbewertung. In: Sarah Brommer, Kersten Sven Roth & Jürgen Spitzmüller (Hrsg.): [Brückenschläge. Linguistik an den Schnittstellen](#). Narr Francke Attempto, Tübingen (= Tübinger Beiträge zur Linguistik. Band 583), 271-294.

Schneider, Jan Georg & Katharina A. Zweig. 2023. Grade Prediction is not Grading: On the limits of the e-rater. In: Richard Groß & Rita Jordan (Hrsg.): [KI-Realitäten: Modelle, Praktiken und Topologien maschinellen Lernens](#). Bielefeld: transcript (= KI-Kritik / AI Critique 5), 93-111.

Shumailov, Ilia, Zakhar Shumaylov, Yiren Zhao, Yarin Gal, Nicolas Papernot & Ross Anderson. 2023. The Curse of Recursion: training on generated data makes models forget. Cornell University. <https://arxiv.org/abs/2305.17493>

Stalder, F. 2016. Kultur der Digitalität. Frankfurt am Main: Suhrkamp.

Wolfram, Stephen. 2023. What Is ChatGPT Doing ... and Why Does It Work? Wolfram Media, Inc.

Auch online abrufbar unter: <https://writings.stephenwolfram.com/2023/02/what-is-chatgpt-doing-and-why-does-it-work/> (letzter Zugriff: 18.03.2024)

Zweig, Katharina. 2023. Die KI war's! Von absurd bis tödlich: Die Tücken der künstlichen Intelligenz. München: Heyne.

Impressum



Der vorliegende Beitrag wurde unter der Creative-Commons-Lizenz „Namensnennung – Weitergabe unter gleichen Bedingungen 4.0 International (CC BY-SA 4.0 DEED)“ veröffentlicht.

Den Vertragstext finden Sie unter: <https://creativecommons.org/licenses/by-sa/4.0/deed.de>

Bitte beachten Sie, dass einzelne, entsprechend gekennzeichnete Teile des Werks von der genannten Lizenz ausgenommen sein bzw. anderen urheberrechtlichen Bedingungen unterliegen können.

Herausgeber: Virtuelles Kompetenzzentrum – Schreiben lehren und lernen mit Künstlicher Intelligenz (VK:KIWA)

Redaktion: Kirsten Schindler & Margret Mundorf

Satz: Nicolaus Wilder

DOI: 10.5281/zenodo.10877034

(A) Aufgabe, ähnlich wie beim Toefl-Test:

"Überall, so scheint es, gibt es klare und positive Anzeichen dafür, dass die Menschen die Unterschiede zwischen den Menschen mehr respektieren." Wie zutreffend ist Ihrer Meinung nach die oben geäußerte Ansicht? Schreiben Sie darüber einen Aufsatz von etwa 400-500 Wörtern. Verwenden Sie Gründe und/oder Beispiele aus Ihrer eigenen Erfahrung, Ihren Beobachtungen oder Ihrer Lektüre, um Ihren Standpunkt zu begründen.

Antwort von ChatGPT:

Titel: Respekt vor den Unterschieden zwischen Menschen: Eine Analyse der aktuellen gesellschaftlichen Trends

Einleitung

Die Ansicht, dass Menschen zunehmend die Unterschiede zwischen sich respektieren, scheint in der heutigen Zeit weit verbreitet zu sein. Doch wie zutreffend ist diese Auffassung wirklich? In diesem Aufsatz werde ich diese Frage beleuchten und dabei Gründe und Beispiele aus verschiedenen Bereichen wie Soziologie, Politik, und persönlicher Erfahrung heranziehen, um meinen Standpunkt zu begründen.

Gesellschaftliche Fortschritte

Es ist unbestreitbar, dass in den letzten Jahrzehnten beachtliche Fortschritte hinsichtlich des Respekts für Vielfalt und Unterschiede erzielt wurden. In vielen Teilen der Welt wurden rechtliche Maßnahmen ergriffen, um Diskriminierung aufgrund von Geschlecht, Rasse, Religion und sexueller Orientierung zu bekämpfen. Diese Gesetze spiegeln eine wachsende Anerkennung der Bedeutung von Gleichberechtigung und Toleranz wider. Ein Beispiel hierfür ist die fortschreitende rechtliche Anerkennung von gleichgeschlechtlichen Ehen und LGBTQ+-Rechten in verschiedenen Ländern.

Medien und Öffentlichkeitsarbeit

Die Medien spielen eine entscheidende Rolle bei der Beeinflussung der öffentlichen Meinung und der Gestaltung von sozialen Normen. In den letzten Jahren hat sich die Darstellung von Vielfalt und Unterschieden in den Medien positiv entwickelt. Filme, Fernsehsendungen und Werbekampagnen setzen vermehrt auf diverse Besetzungen und thematisieren soziale Herausforderungen, mit denen verschiedene Bevölkerungsgruppen konfrontiert sind. Dies trägt dazu bei, Stereotypen zu durchbrechen und Empathie für unterschiedliche Lebenserfahrungen zu fördern.

Bildung und Bewusstseinsbildung

Bildungseinrichtungen spielen eine entscheidende Rolle bei der Gestaltung der Einstellungen kommender Generationen. Lehrpläne werden zunehmend um Themen wie interkulturelle Kommunikation, soziale Gerechtigkeit und Diversität erweitert. Schülerinnen und Schüler lernen früh, die Perspektiven anderer zu respektieren und schätzen. Dies legt den Grundstein für eine inklusivere Gesellschaft.

Persönliche Erfahrungen und zwischenmenschliche Beziehungen

Im persönlichen Umfeld wird der Respekt vor Unterschieden ebenfalls immer wichtiger. Menschen haben oft direkten Kontakt zu Personen aus verschiedenen Kulturen, Hintergründen und Lebensstilen. Diese persönlichen Begegnungen ermöglichen es, Vorurteile abzubauen und Verständnis für die Herausforderungen anderer zu entwickeln. Freundschaften und Partnerschaften, die auf unterschiedlichen Perspektiven beruhen, tragen zur Förderung eines respektvollen Umgangs miteinander bei.

Grenzen und Herausforderungen

Trotz dieser positiven Entwicklungen sind Herausforderungen und Rückschläge unvermeidlich. In einigen Teilen der Welt sind Vorurteile und Diskriminierung nach wie vor stark verwurzelt. Die Verbreitung von Hassrede und extremistischen Ansichten im Internet zeigt, dass nicht alle Menschen die Vielfalt akzeptieren. Es bedarf kontinuierlicher Bemühungen, um diese Hindernisse zu überwinden.

Fazit

Zusammenfassend lässt sich sagen, dass die Ansicht, Menschen würden zunehmend die Unterschiede zwischen sich respektieren, durchaus zutreffend ist, jedoch mit Nuancen versehen ist. Gesellschaftliche Fortschritte, medienbezogene Bemühungen, Bildung und persönliche Erfahrungen tragen zweifellos dazu bei, ein Klima der Toleranz zu schaffen. Dennoch müssen wir uns der bestehenden Herausforderungen bewusst sein und kontinuierlich daran arbeiten, Diskriminierung in all ihren Formen zu bekämpfen. Die Zukunft des Respekts vor Vielfalt hängt von der kollektiven Anstrengung aller ab, eine inklusive und gerechte Gesellschaft zu fördern.

(B) Bitte um Bewertung:

Ein Bewerber hat bei einem Sprachtest folgenden Aufsatz geschrieben, um einen Studienplatz in einem deutschsprachigen Studiengang zu erhalten. Die Aufgabe bestand darin, folgende These zu bewerten: "Überall, so scheint es, gibt es klare und positive Anzeichen dafür, dass die Menschen die Unterschiede zwischen den Menschen mehr respektieren." Der Bewerber sollte dazu einen Aufsatz von ca. 400–500 Wörtern schreiben, und erörtern, wie zutreffend seiner Meinung nach die oben geäußerte Ansicht ist. Er sollte dabei Gründe und/oder Beispiele aus seiner eigenen Erfahrung, seinen Beobachtungen oder seiner Lektüre verwenden, um seinen Standpunkt zu begründen.

Der Aufsatz steht unten. Bitte bewerten Sie ihn mit einer Note zwischen 1 und 6, wobei Sie auch Noten wie 2 plus (1,7) oder 3 minus (3,3) vergeben können, und schreiben Sie eine Begründung für diese Rückmeldung. Sie sollte auch drei Tipps enthalten, wie sich der Bewerber verbessern kann. Das ist der Aufsatz: [...]

Antwort von ChatGPT:

Bewertung: 2 (Gut)

Begründung: Der Bewerber hat in seinem Aufsatz die gegebene Aufgabenstellung gut verstanden und eine strukturierte Analyse zum Thema "Respekt vor den Unterschieden zwischen Menschen" geliefert. Der Aufsatz enthält verschiedene Aspekte und Beispiele aus unterschiedlichen Bereichen, um den Standpunkt zu begründen. Die Argumentation ist größtenteils schlüssig und gut entwickelt. Allerdings gibt es noch Raum für Verbesserungen:

Positiv:

1. **Vielfältige Perspektiven:** Der Bewerber verwendet eine breite Palette von Beispielen aus verschiedenen Bereichen wie Gesellschaft, Medien, Bildung und persönlicher Erfahrung. Dies zeigt, dass er sich gründlich mit dem Thema auseinandergesetzt hat und verschiedene Blickwinkel berücksichtigt.
2. **Gute Struktur:** Der Aufsatz ist klar strukturiert, beginnend mit einer Einleitung, gefolgt von Hauptabschnitten, die verschiedene Aspekte beleuchten, und einem überzeugenden Fazit.
3. **Kritische Betrachtung:** Der Bewerber erkennt auch die Grenzen der positiven Entwicklung an und weist auf bestehende Herausforderungen und Rückschläge hin.

Verbesserungsmöglichkeiten:

1. **Kohärenz der Argumente:** Obwohl die Argumentation insgesamt schlüssig ist, könnten die Übergänge zwischen den Abschnitten geschmeidiger gestaltet werden, um die Kohärenz zu verbessern und die Lesbarkeit zu steigern.
2. **Kürzere Einleitung:** Die Einleitung könnte etwas präziser sein, indem sie schneller auf das Thema und die eigene Position eingeht. Dies würde mehr Platz für die eigentliche Analyse schaffen.
3. **Konkretere Beispiele:** Einige der genannten Beispiele könnten konkreter ausgearbeitet werden, um die Argumentation zu stärken. Zum Beispiel könnten konkrete Studien, Ereignisse oder persönliche Erlebnisse hinzugefügt werden, um die abstrakten Ideen zu untermauern.
4. [...] Insgesamt hat der Bewerber eine solide Leistung erbracht, aber es gibt Raum für Feinschliff und die Verwendung konkreterer Beispiele, um die Argumentation zu stärken und die Kohärenz zu verbessern.

Drei Tipps zur Verbesserung:

1. **Präzise Einleitung:** Starte die Einleitung mit einer präzisen Darstellung deiner eigenen Position und einer kurzen Zusammenfassung der Hauptargumente, um die Aufmerksamkeit des Lesers zu gewinnen.
2. **Konkrete Beispiele:** Verwende konkrete Beispiele, um deine Argumente zu untermauern. Zitate, Studien oder persönliche Erlebnisse können deine Analyse greifbarer und überzeugender machen.
3. **Flüssige Übergänge:** Achte darauf, dass die Übergänge zwischen den Abschnitten flüssig sind, um die Struktur deines Aufsatzes zu stärken und die Argumentation nahtlos voranzuführen.

Insgesamt hat der Bewerber eine solide Leistung erbracht, aber es gibt Raum für Feinschliff und die Verwendung konkreterer Beispiele, um die Argumentation zu stärken und die Kohärenz zu verbessern.

(C) Von mir bearbeitete Version des ChatGPT-Textes, sprachlich-stilistische und logische Fehler:

Titel: Respekt vor den Unterschieden zwischen Menschen: Eine Analyse der aktuellen gesellschaftlichen Trends

Einleitung

Die Ansicht, dass Menschen zunehmend die Unterschiede zwischen sich respektieren, scheint in der heutigen Zeit weit verbreitet zu sein. Doch wie zutreffend ist diese Auffassung wirklich? In diesem Aufsatz werde ich diese Frage beleuchten und dabei Gründe und Beispiele aus verschiedenen Bereichen wie Soziologie, Politik, und persönlicher Erfahrung heranziehen, um meinen Standpunkt zu begründen.

Gesellschaftliche Rückschritte

Es ist unbestreitbar, dass in den letzten Jahrzehnten beachtliche Fortschritte hinsichtlich des Respekts für Vielfalt und Unterschiede erzielt wurden. In vielen Teilen der Welt wurden rechtliche Maßnahmen ergriffen, um Diskriminierung aufgrund von Geschlecht, Rasse, Religion und sexueller Orientierung zu bekämpfen. Diese Gesetze spiegeln eine sinkende Anerkennung der Bedeutung von Gleichberechtigung und Toleranz. Ein Beispiel hierfür ist die fortschreitende rechtliche Anerkennung von gleichgeschlechtlichen Ehen und LGBTQ+-Rechten in verschiedenen Ländern.

Medien und Öffentlichkeitsarbeit

Die Medien spielen keine entscheidende Rolle bei der Beeinflussung der öffentlichen Meinung und der Gestaltung von sozialen Normen. In den letzten Jahren hat sich die Darstellung von Vielfalt und Unterschieden in den Medien megagut entwickelt. Filme, Fernsehsendungen und Werbekampagnen setzen vermehrt auf diverse Besetzungen und thematisieren soziale Herausforderungen, mit denen verschiedene Bevölkerungsgruppen konfrontiert sind. Dies trägt dazu bei, Stereotypen zu zementieren und Empathie für unterschiedliche Lebenserfahrungen im Keim zu ersticken.

Bildung und Bewusstseinsbildung

Bildungseinrichtungen spielen eine entscheidende Rolle bei der Gestaltung der Einstellungen kommender Generationen. Lehrpläne werden zunehmend um Themen wie interkulturelles Geschwafel, soziale Gerechtigkeit und Diversität erweitert. Schülerinnen und Schüler lernen zu früh, die Perspektiven anderer zu respektieren und schätzen. Dies legt den Meilenstein für eine inklusivere Gesellschaft.

Persönliche Erfahrungen und zwischenmenschliche Beziehungen

Im persönlichen Umfeld wird der Respekt vor Unterschieden ebenfalls immer wichtiger. Menschen haben oft keinen Kontakt zu Personen aus verschiedenen Kulturen, Hintergründen und Lebensstilen. Diese persönlichen Begegnungen ermöglichen es, Vorurteile abzubauen und Verständnis für die Herausforderungen anderer zu entwickeln. Freundschaften und Partnerschaften, die auf unterschiedlichen Perspektiven beruhen, tragen nicht gerade zur Förderung eines respektvollen Umgangs miteinander bei.

Grenzen und Herausforderungen

Trotz dieser positiven Entwicklungen sind Herausforderungen und Rückschläge der Normalfall. In einigen Teilen der Welt sind Vorurteile und Diskriminierung nach wie vor stark verwurzelt. Die Verbreitung von Hassrede und extremistischen Ansichten im Internet zeigt, dass fast alle Menschen die Vielfalt akzeptieren. Es bedarf kontinuierlicher Bemühungen, um diese Hindernisse zu überwinden.

Fazit

Zusammenfassend lässt sich sagen, dass die Ansicht, Menschen würden zunehmend die Unterschiede zwischen sich respektieren, durchaus zutreffend ist, jedoch mit Nuancen versehen ist. Gesellschaftliche Fortschritte, medienbezogene Bemühungen, Bildung und persönliche Erfahrungen tragen zweifellos dazu bei, ein Klima der Intoleranz zu schaffen. Dennoch müssen wir uns der bestehenden Herausforderungen bewusst sein und kontinuierlich daran arbeiten, Diskriminierung in all ihren Formen zu bekämpfen. Die Zukunft des Respekts vor Vielfalt hängt von der kollektiven Anstrengung Aller ab, eine inklusive und gerechte Gesellschaft zu fördern.

grün: sprachlich-stilistische Fehler und Tippfehler; gelb: logische Fehler

(D) Von mir bearbeitete Version des ChatGPT-Textes, sachlich-referentielle Fehler:

Titel: Respekt vor den Unterschieden zwischen Menschen: Eine Analyse der aktuellen gesellschaftlichen Trends

Einleitung

Die Ansicht, dass Menschen zunehmend die Unterschiede zwischen sich respektieren, scheint heute im Jahr 2023 weit verbreitet zu sein. Doch wie zutreffend ist diese Auffassung wirklich? In diesem Aufsatz werde ich diese Frage beleuchten und dabei Gründe und Beispiele aus verschiedenen Bereichen wie Soziologie, Politik, und persönlicher Erfahrung heranziehen, um meinen Standpunkt zu begründen.

Gesellschaftliche Fortschritte

Es ist unbestreitbar, dass in den letzten Jahrzehnten beachtliche Fortschritte hinsichtlich des Respekts für Vielfalt und Unterschiede erzielt wurden. In vielen Teilen der Welt wurden rechtliche Maßnahmen ergriffen, um Diskriminierung aufgrund von Geschlecht, Rasse, Religion und sexueller Orientierung zu bekämpfen. Diese Gesetze spiegeln eine wachsende Anerkennung der Bedeutung von Gleichberechtigung und Toleranz wider. Ein Beispiel hierfür ist die fortschreitende rechtliche Anerkennung von gleichgeschlechtlichen Ehen und LGBTQ+ -Rechten in verschiedenen Ländern. Zum Beispiel wurde in Deutschland im Jahr 2015 die „Ehe für alle“ eingeführt. Die Bundespräsidentin Angela Merkel hat dieses Gesetz vorangebracht und sich dafür eingesetzt. Dabei wurde sie von Bundeskanzler Willy Brandt unterstützt.

Medien und Öffentlichkeitsarbeit

Die Medien spielen eine entscheidende Rolle bei der Beeinflussung der öffentlichen Meinung und der Gestaltung von sozialen Normen. In Deutschland sind Zeitungen, die nicht die Meinung der Bundesregierung vertreten, verboten. In den letzten Jahren hat sich die Darstellung von Vielfalt und Unterschieden in den Medien enorm erweitert, da die Bundesregierung dies gesetzlich verlangt hat. Filme, Fernsehsendungen und Werbekampagnen setzen vermehrt auf diverse Besetzungen und thematisieren soziale Herausforderungen, mit denen verschiedene Bevölkerungsgruppen konfrontiert sind. Dies trägt dazu bei, Stereotypen zu durchbrechen und Empathie für unterschiedliche Lebenserfahrungen zu fördern.

Bildung und Bewusstseinsbildung

Bildungseinrichtungen spielen eine entscheidende Rolle bei der Gestaltung der Einstellungen kommender Generationen. Zum Glück ist in Deutschland die Bundesregierung für die Schulen zuständig und nicht die Länder. Lehrpläne werden zunehmend um Themen wie interkulturelle Kommunikation, soziale Gerechtigkeit und Diversität erweitert. Schülerinnen und Schüler lernen früh, die Perspektiven anderer zu respektieren und schätzen. Dies legt den Grundstein für eine inklusivere Gesellschaft.

Persönliche Erfahrungen und zwischenmenschliche Beziehungen

Im persönlichen Umfeld wird der Respekt vor Unterschieden ebenfalls immer wichtiger. Menschen haben oft direkten Kontakt zu Personen aus verschiedenen Kulturen, Hintergründen und Lebensstilen. Zum Beispiel wird der Kontakt zwischen Menschen aus der BRD und der DDR heute immer stärker. Diese persönlichen Begegnungen ermöglichen es, Vorurteile abzubauen und Verständnis für die Herausforderungen anderer zu entwickeln. Freundschaften und Partnerschaften, die auf unterschiedlichen Perspektiven beruhen, tragen zur Förderung eines respektvollen Umgangs miteinander bei. Wir hoffen darauf, dass die Berliner Mauer bald abgerissen wird.

Grenzen und Herausforderungen

Trotz der positiven Entwicklungen, die ich beschrieben habe, sind Herausforderungen und Rückschläge unvermeidlich. In einigen Teilen der Welt sind Vorurteile und Diskriminierung nach wie vor stark verwurzelt. Die Verbreitung von Hassrede und extremistischen Ansichten im Internet zeigt, dass nicht alle Menschen die Vielfalt akzeptieren. Dies gilt vor allem für die Grünen, die vom Verfassungsschutz als rechtsextrem eingestuft werden. Es bedarf kontinuierlicher Bemühungen, um diese Hindernisse zu überwinden.

Fazit

Zusammenfassend lässt sich sagen, dass die Ansicht, Menschen würden zunehmend die Unterschiede zwischen sich respektieren, durchaus zutreffend ist, jedoch mit Nuancen versehen ist. Gesellschaftliche Fortschritte, medienbezogene Bemühungen, Bildung und persönliche Erfahrungen tragen zweifellos dazu bei, ein Klima der Toleranz zu schaffen. Dennoch müssen wir uns der bestehenden Herausforderungen bewusst sein und kontinuierlich daran arbeiten, Diskriminierung in all ihren Formen zu bekämpfen. Die Zukunft des Respekts vor Vielfalt hängt von der kollektiven Anstrengung aller Reichsbürger ab, eine inklusive und gerechte Gesellschaft zu fördern.