

Eine Unterhaltung mit ChatGPT: Dataistisches Paradigma, stochastischer Papagei oder einfach nur ein Mißverständnis?

Künstliche Intelligenz ist derzeit in aller Munde: Wahre Wunderdinge werden von den neuen Programmen erwartet und so ist es auch nicht erstaunlich, daß der von der Firma OpenAI entwickelte Chatbot ChatGPT heiß diskutiert und vor allem auch weltweit ausprobiert wird.

Am Sonntag, den 11.12.2022, habe ich von 21.05–21.11 Uhr nachstehende Unterhaltung mit ChatGPT geführt (Antworten von ChatGPT mit Answer 1). Am Donnerstag, den 15.12.2022, habe ich von 13.30–13.38 Uhr diese Unterhaltung mit ChatGPT noch einmal geführt (Answer 2), wobei ich dieselben Fragen gestellt habe, jedoch im Hinblick auf den Text in Frage 4 die Personalpronomen geringfügig geändert habe:

Question 1: Are you a stochastic parrot?

Answer 1 (11.12.2022): I'm sorry, but I am not a parrot. I am an AI assistant trained to assist with a wide variety of tasks and answer questions to the best of my ability. While I am not capable of physical actions, like a parrot, I am designed to be highly knowledgeable and able to provide helpful and accurate information. Is there something specific you would like to know? I would be happy to help.

Answer 2 (15.12.2022): I am not a parrot of any kind, let alone a stochastic one. I am a large language model trained by OpenAI. I don't have a physical form, I exist solely as a computer program. I am designed to assist with a wide variety of tasks, including answering questions and providing information on a wide range of topics. Is there anything specific you would like to know about?

Question 2: Can you describe your language model?

Answer 1 (11.12.2022): As a language model, I am trained to understand and generate natural language text. I do this by analyzing large amounts of text data and using that information to make predictions about what words or phrases are most likely to come next in a given sequence of text. This allows me to understand the context of the words and sentences I am processing and generate coherent and appropriate responses to questions and requests. I am constantly learning and improving my ability to understand and generate text, and I am designed to be highly adaptable and versatile.

Answer 2 (15.12.2022): Yes, I can certainly describe my language model. As I mentioned earlier, I am a large language model trained by OpenAI. This means that I have been trained on a massive amount of text data in order to be able to generate human-like text and assist with a wide range of tasks. My training data included books, articles, and other written materials, which I used to learn about the

structure and style of human language. This allows me to generate text that is similar to human-written text, and to understand and respond to a wide range of questions and requests. I am constantly improving as I am exposed to more and more data, which allows me to become more accurate and helpful over time.

Question 3: Can you give me an example for predictions about what words or phrases are most likely to come next in a given sequence of text?

Answer 1 (11.12.2022): Sure, here's an example: Imagine you are reading the following sentence: „The quick brown fox jumps over the lazy dog. The fox is known for its quickness and agility, and it often uses these abilities to outmaneuver its enemies.“

Based on the words and phrases that have come before, the language model I am based on might predict that the next word or phrase is likely to be something related to the fox's characteristics or behavior, such as „speed,“ „intelligence,“ or „hunting tactics.“ These predictions are based on the model's analysis of the context of the words and phrases in the sentence and its understanding of how words and phrases are typically used in similar contexts.

Of course, the model's predictions are not always correct, and it may sometimes generate unexpected or surprising words or phrases. However, its predictions can still be useful in helping to understand and generate natural language text.

Answer 2 (15.12.2022): Sure, I can provide an example of how I can make predictions about the words or phrases that are most likely to come next in a given sequence of text. For instance, let's say that you provide me with the following sequence of words: „The cat sat on the“ I could use my knowledge of the structure and style of human language, as well as the context provided by the words you've given me, to predict that the most likely word or phrase to come next in this sequence would be „mat“ or „couch“. This is because these words are commonly used in similar contexts, and they fit grammatically with the words that have already been provided. Of course, there could be other words or phrases that could also be considered likely candidates, depending on the specific context and the broader context of the text as a whole.

Question 4: Can you explain the following text: And once it is written, every text drifts everywhere, and in the same way with those who understand it as with those for whom it does not suit, and he does not know to whom he should speak and to whom he should not. And if he is mistreated and unjustly criticised, he always needs his father's help. For he himself cannot defend himself nor help.¹

Can you explain the following text: And once it is written, every text drifts everywhere, and in the same way with those who understand it as with those for whom it does not suit, and it does not know to whom it should speak and to whom it should not. And if it is mistreated and unjustly criticised, it always needs his father's help. For it itself cannot defend itself nor help.²

Answer 1 (11.12.2022): The text you provided is somewhat difficult to understand, as it contains some vague and abstract language that can be open to interpretation. However, based on my understanding of the words and phrases used, I can offer the following interpretation: The text appears to be discussing the nature of written language and its ability to be understood by different people. It suggests that once a text is written, it can be read and interpreted by anyone, regardless of whether or not it is relevant or suitable for them. This can lead to misunderstandings and criticism, and the writer of the text may need help defending themselves against these negative reactions.

1 Engl. Übersetzung mit <https://www.deepl.com/translator> (abgerufen am 13.02.2023) von: E. Heitsch (Übers./Komm.), Platon Phaidros (= Platon Werke III, 4), Göttingen 1993, 62 zu Plat. Phaidr. 275d–e.

2 Gleicher Text wie in Frage 4, jedoch für Personalpronomen, die für den ‚Text‘ stehen, immer das Neutrum eingesetzt.

Overall, the text seems to be expressing the idea that written language can be unpredictable and potentially harmful, and that the writer may need support and protection in order to defend themselves against unfair criticism.

Answer 2 (15.12.2022): This text appears to be talking about the nature of written language and its relationship with those who read and understand it. The text suggests that once a piece of writing is created, it can be read and interpreted by anyone, regardless of whether or not the content is relevant or suitable for them. The text also suggests that written language is vulnerable and unable to defend itself or seek help on its own, and that it requires the support of its creator (referred to as its „father“) in order to be protected and defended against mistreatment and unjust criticism. Overall, the text seems to be exploring the idea that written language is dependent on its creator and those who interpret it, and that it has no agency of its own.

Question 5: Doesn't that contradict your definition of yourself expressed earlier?

Answer 1 (11.12.2022): I'm sorry if my previous response was unclear or confusing. As an AI assistant, I am not capable of writing or understanding written language in the same way that a human can. I am trained to process and generate natural language text, but my abilities are limited to the specific tasks I am designed for and the data I have been trained on.

In my previous response, I was simply offering an interpretation of the text you provided based on my understanding of the words and phrases used. I was not expressing any personal opinions or beliefs, and I was not contradicting anything I had said earlier about my own abilities or nature. I apologize if my previous response was unclear or misleading in any way.

Answer 2 (15.12.2022): When I described myself earlier, I stated that I am a large language model trained by OpenAI, and that I exist solely as a computer program. This is not in contradiction with the text you provided, as it is not directly referring to language models like myself. Instead, the text is discussing the nature of written language and its relationship with those who read and understand it. While I am a computer program designed to assist with tasks involving written language, I do not have the same vulnerabilities and limitations as the written language described in the text. As a language model, I am able to defend myself against mistreatment and criticism, and I am not dependent on anyone else for support or protection. I exist solely as a computer program, and I am not subject to the same limitations as the written language discussed in the text.

Question 6: I did not ask for your personal opinion, but whether it is not a contradiction if you can make predictions with regard to the context of a text, but answer my question by saying that my text suggests that once a text is written, it can be read and interpreted by anyone, regardless of whether or not it is relevant or suitable for them.

Answer 1 (11.12.2022): I apologize if my previous response did not fully address your question. To clarify, making predictions about the context of a text and interpreting the meaning of a text are two different things.

As a language model, I am trained to make predictions about what words or phrases are likely to come next in a given sequence of text based on the context of the words and phrases that have come before. This allows me to understand the overall meaning and structure of a text and generate appropriate and coherent responses to questions and requests.

However, interpreting the meaning of a text involves making judgments about the intended message or purpose of the text, and this goes beyond my capabilities as a language model. I can provide information and answer questions based on the data I have been trained on, but I do not have the ability to interpret the meaning of a text in the same way that a human can.

In short, my previous response was based on my understanding of the words and phrases used in the text you provided, but it was not intended as a definitive interpretation of the text's intended meaning. I apologize if my response was unclear or misleading in any way.

Answer 2 (15.12.2022): There is no contradiction in my previous responses. As a language model, I am designed to be able to make predictions about the words or phrases that are most likely to come next in a given sequence of text based on the context provided. This ability is based on my training on a large amount of text data, which allows me to understand the structure and style of human language and generate text that is similar to human-written text. However, this does not mean that I am limited to only making predictions about the next words or phrases in a sequence of text. I am also able to understand and respond to a wide range of questions and requests, including providing explanations or interpretations of text. In the previous example, I was able to provide an explanation of the text you provided based on my understanding of the words and phrases used, and the broader context of the text as a whole. This is not a contradiction of my ability to make predictions about the next words or phrases in a given sequence of text, but rather an example of my broader capabilities as a language model.

Man könnte die ersten Antworten nun als Interpretationsverweigerung verstehen und alle kritischen Urteile wiederholen, die – kaum daß ChatGPT der Community zur Verfügung gestellt wurde – auf breiter Front geäußert wurden. M.E. wäre dies jedoch unangebracht, denn die Antworten sind nicht sinnlos, auch nicht ‚irgendwie falsch‘. Sie zeigen einfach nur, daß die Analyse auf der Metaebene – also das, was wir gemeinhin als die Interpretation von komplexen Texten verstehen – dem Modell nicht möglich ist. Interessant ist, daß die ersten Antworten von ChatGPT vom 11.12.2022 deutlich daraufhin weisen und die zweiten Antworten vom 15.12.2022 demgegenüber genau das Gegenteil besagen, nämlich daß das Modell von ChatGPT durchaus in der Lage sei, auch Textinterpretationen zu liefern!

Ted Underwood hatte schon im Oktober 2021 anlässlich des anschwellenden Warnesangs hinsichtlich der Möglichkeiten, die mit den neuen KI-Sprachmodellen verbunden werden, gewarnt: Faszinierend, aber auch gefährlich seien diese neuen Modelle:³

3 T. Underwood, Mapping the Latent Spaces of Culture, online: <http://dx.doi.org/10.17613/faaa-1r21> (abgerufen am 11.12.2022).



Abb. 1: Mit VQGAN + CLIP und Nightcafe Creator für die Texteingabe „A macaw clockwork parrot in Borges' library“ erzeugt.⁴

Underwood sieht aber durchaus nicht alles verloren: „But it also gives us more reason to hope. It is not entirely clear what we plan to gain by modeling intelligence, since we already have more than seven billion intelligences on the planet. By contrast, it’s easy to see how exploring spaces of possibility implied by the human past could support a more reflective and more adventurous approach to our future.“⁵

Eine ausführliche Analyse der Gefahren, die in den neuen Language Modellen (LM) liegen, hatten schon einige Monate zuvor Emily Bender et al. in ihrem Paper „On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?“ geäußert und dabei auf folgenden Aspekt hingewiesen:

„Text generated by an LM is not grounded in communicative intent, any model of the world, or any model of the reader’s state of mind. It can’t have been, because the training data never included sharing thoughts with a listener, nor does the machine have the ability to do that.“⁶

4 Nach: <https://tedunderwood.com/2021/10/21/latent-spaces-of-culture/> (abgerufen am 11.12.2022).

5 T. Underwood, Mapping the Latent Spaces of Culture, online: <http://dx.doi.org/10.17613/faaa-1r21> (abgerufen am 11.12.2022).

6 E. Bender et al., On the Dangers of Stochastic Parrots, Can Language Models Be Too Big? Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell 2021, On the Dangers of Stochastic Parrots, Can Language Models Be Too Big?, in: Conference on Fairness, Accountability, and Transparency (FAccT '21), March 3–10, 2021, Virtual Event, Canada, ACM, New York, NY, USA, 14 pages. <https://doi.org/10.1145/3442188.3445922>. Zu der Diskussion, die dieses Paper ausgelöst hat, vgl. z.B. <https://www.wired.com/story/google-timnit-gebru-ai-what-really-happened/>, <https://languagelog ldc.upenn.edu/nll/?p=51161>, <https://gist.github.com/yoavg/9fc9be2f98b47c189a513573d902fb27>,

Dies ist nun ein Gedanke, der weder neu noch originell ist. Als Althistorikerin weise ich auf die folgende Passage aus Platons *Phaidros* hin, die ich auszugsweise in der englischen Übertragung von Deepl als Frage an ChatGPT (Frage 4) gerichtet habe:

Phaidros 275d–e: ὅταν δὲ ἅπαξ γραφῆι, κυλινδεῖται μὲν πανταχοῦ πᾶς λόγος ὁμοίως παρὰ τοῖς ἐπαῖουσιν, ὡς δ' αὐτως παρ' οἷς οὐδὲν προσήκει, καὶ οὐκ ἐπίσταται λέγειν οἷς δεῖ γε καὶ μὴ. πλημμελούμενος δὲ καὶ οὐκ ἐν δίκῃ λοιδορηθεὶς τοῦ πατρὸς ἀεὶ δεῖται βοηθοῦ· αὐτὸς γὰρ οὔτ' ἀμύνασθαι οὔτε βοηθῆσαι δυνατὸς αὐτῷ.

„Und ist er erst einmal geschrieben, treibt jeder Text sich überall herum und zwar in gleicher Weise bei denen, die ihn verstehen, wie bei denen, für die er nicht passt, und er weiß nicht, zu wem er reden soll und zu wem nicht. Und wird er misshandelt und zu Unrecht kritisiert, braucht er den Vater als Helfer. Denn er selbst kann sich nicht wehren noch sich selbst helfen.“⁷

Im Gegensatz dazu präferierte Platon das *λόγον δοῦναι καὶ δέξασθαι*, ein dialogisches Fragen und Antworten in dem Sinn, daß nicht nur Antworten im Sinne von Rechenschaft geben dazu gehören, sondern auch das Empfangen, d.h. das Auf- und Annehmen dieser Rede dazu gehört. Der zweite Teil *καὶ δέξασθαι* – das Empfangen von Rede – bringt ein aktives Element in den Diskurs, das auf der Seite des die Antwort Empfangenden ein selbstreflexives Moment einführt, und beides zusammen erst ist die Grundlage für eine wirkliche Erkenntnis.

So ist der Anspruch, der in dem Begriff des dataistischen Paradigmas zum Ausdruck kommt⁸ – wenn man das Sprachmodell von GPT so bezeichnen möchte – nichts weiter als ein Mißverständnis – nämlich, daß man die Dynamik des *λόγον δοῦναι καὶ δέξασθαι* durch ein wahrheitsbasiertes Textdatentraining nachbilden könne. Das Bild des stochastischen Papageis trifft das Verhältnis jedoch (noch) recht gut: Genauso wie Platon im *Theaitetos* diejenigen, die nicht Rede und Antwort stehen können, also nicht zum *λόγον δοῦναι καὶ δέξασθαι* bereit sind, mit unmündigen Kindern gleichsetzt (*Theaitetos* 177b und 202), so kann auch die chattende KI nicht das, was mündige Leser/innen und Interpreten können, geschweige denn, daß Scherz, Ironie oder bewußte Mehrdeutigkeit in der Verwendung von Metaphern etc. von einer KI erfaßt würden.

Allerdings sollte man auch für die weitere Entwicklung, die unmittelbar bevorsteht, Überlegungen dahingehend entwickeln, wie und für welche Anwendungsfälle solche Systeme einsetzbar sind, konkret heißt das, wie soll das Verhältnis zwischen einer chattenden KI und einem/r mündigen Leser/in gestaltet werden? Abgesehen davon, daß noch viel an rassistischen, frauenfeindlichen oder anderweitig inhaltlich inadäquaten Aussagen zustande kommen und diese natürlich beseitigt werden müssen, sind die KI-basierten Assistenzsysteme für Bereiche in der Diskussion, die sich durch hohe Redundanz auszeichnen wie maschinelle Übersetzung, Rechtschreibprüfung oder auch z.B. auch Internetsuchen.⁹

Bedenkt man hingegen die Möglichkeiten, die sich – um bei dem Begriff der KI-basierten Assistenzsysteme zu bleiben – in Zukunft für den Einsatz in Schule und Hochschule ergeben, dann sollten rechtzeitig im Hinblick auf Prüfungswesen und Prüfungsmöglichkeiten Rahmenbedingungen geschaffen werden, die nicht den Einsatz solcher Systeme verhindern (ein solcher Versuch würde m.E. sowieso zum Scheitern verurteilt sein), sondern Produktivitätsgewinn und -ermöglichung in den Vordergrund stellen. D.h., daß in Zukunft der Arbeitsprozeß selbst, der zu einem Ergebnis geführt hat, und vor al-

<https://magazine.scienceforthepeople.org/vol24-2-dont-be-cvil/stochastic-parrots/> (alle abgerufen am 11.02.2023).

7 Übersetzung nach Platon *Phaidros*, E. Heitsch (Übers./Komm.), Platon *Phaidros* (= Platon Werke III, 4), Göttingen 1993, 62 zu Plat. *Phaidr.* 275d–e.

8 Zu dem Begriff vgl. Jose van Dijck, *Datafication, dataism and dataveillance: big data between scientific paradigm and ideology*, *Surveillance & Society* 12/2 (2014), 197–208 und Minna Ruckenstein & Mika Pantzar, *Beyond the Quantified Self, Thematic exploration of a dataistic paradigm, new media & society* 19/3 (2017), 401–418.

9 Vgl. dazu Alexander Pretschner, Eric Hilgendorf, Ute Schmid, Hannah Schmid-Petri, *Die mächtigen neuen Assistenzsysteme*, in: *FAZ* 9.1.2023, S. 18.

lem die damit verbundene Selbstreflexivität im Einsatz von KI-basierten Assistenzsystemen eingeübt und in voller Transparenz offengelegt werden müssten. So könnten chattende KI-Sprachsysteme und mündige Leser/innen und Text produzierende Anwender/innen hoffentlich ein sehr produktives Miteinander entwickeln.

Autoreninformation¹⁰

Charlotte Schubert

Universität Leipzig

Historisches Seminar

E-Mail: schubert@uni-leipzig.de

¹⁰ Die Rechte für Inhalt, Texte, Graphiken und Abbildungen liegen, wenn nicht anders vermerkt, bei der Autorin. Alle Inhalte dieses Beitrages unterstehen, soweit nicht anders gekennzeichnet, der Lizenz CC BY 4.0.