

Protokoll 22

Roberto Simanowski

Die Praxis der kollektiven Autorschaft im Geiste der Statistik

Was mich von Anfang an am meisten an ChatGPT interessierte, war die Frage: Wer eigentlich spricht? Denn ein bisschen mutet so eine Textmaschine ja an wie der „Tod des Autors 2.0“. Wobei es natürlich auf die 2 ankommt, weil bei Barthes der Tod nicht so gemeint war, dass es keine/n konkrete/n Autor/in für einen Text gebe, sondern diese/r lediglich nicht Herr/in seiner oder ihrer Worte ist. Aber mensch identifizierte sich doch immerhin mit dem, was mensch sagte. Denn mensch sagte, was gesagt wurde, und nicht etwa etwas anderes; weil mensch Schnittpunkt eben dieser Diskurse war und nicht etwa jener. Mensch hatte seine Fremdbestimmung verinnerlicht, was schlimmer klingt als es ist. Schlimmer ist das Gegenteil: Die Äußerung ohne jegliche Identifizierung mit dem Gesagten. Genau das ist die Situation von ChatGPT.

ChatGPT produziert Texte ohne Intention und Identifikation, weil es nicht denkt, wenn es schreibt, sondern rechnet. Es erzeugt Wortfolgen nach dem Wahrscheinlichkeitsprinzip; was immer im Pool der Daten, an denen es trainiert wird, die Mehrheit auf seiner Seite hat, bestimmt, wie es weiter geht. Bestimmt also auch, was von strittigen Themen wie Waffenbesitz, Abtreibung, Gleichbehandlung, Meinungsfreiheit, Individualismus, soziale Gerechtigkeit, ziviler Ungehorsam usw. zu halten ist bzw. von ChatGPT als Auskunft dazu ausgegeben wird. Die Botschaft des Mediums ChatGPT – und um Medien handelt es sich wohl bei Sprach-KIs – liegt diesmal auch im Inhalt. ChatGPT verändert die Situation des Menschen doppelt: durch die Automatisierung kognitiver Prozesse und durch die Standardisierung des Kommunikats.

Dass Sprach-KIs wie ChatGPT die Sicht ihrer Nutzer auf die Welt beeinflussen, ist in der KI-Forschung ein Allgemeinplatz.¹ Deswegen all die Forderungen nach „decolonial computing“ (das KI nicht nur an den Daten weißer Männer aus dem globalen Norden trainiert) und „algorithmic reparation“ (die die Daten jener

1 Maurice Jakesch u.a., Co-Writing with Opinionated Language Models Affects Users Views, CHI '23: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, New York 19.4.2023 (<https://dl.acm.org/doi/10.1145/3544548.3581196>)

Gruppen favorisiert, die auch im globalen Norden bisher zu wenig zu Wort kamen).² Allerdings weiß momentan niemand so richtig, wie das „value alignment“ einer global operierenden KI gerecht praktiziert werden kann.

Der Grund ist keineswegs ein technisches Problem, denn auf der technischen Ebene lässt sich die Weltanschauung der KI durch eine entsprechende Filterung der Trainingsdaten und mehr noch durch das Fine-Tuning im Nachgang³ recht zielführend ausrichten. Aber welche Werte sollten es sein? Die westlichen, um sie per KI weltweit – und mit kolonialistischem Effekt – zu exportieren?

Der technologische Entwicklungsstand konfrontiert uns mit einem philosophischen Dilemma: Wir wissen nicht, mit welchen Werten wir die Sprach-KI ausrichten sollen, und wissen doch zugleich, dass wir sie nicht nicht ausrichten können. Unterließen wir es, wäre ChatGPT das Sprachrohr der Mehrheit seiner Daten, was wohl nicht einmal dann wünschenswert wäre, wenn sich diese Mehrheit proportional zur Mehrheit der Menschheit verhielte.

Mit diesen Bedenken im Kopf war meine erste Frage an ChatGPT, am 12. Februar 2023: „Who is the author of a text written by ChatGPT?“

ChatGPT: *The author of a text written by ChatGPT is the person who inputted the prompts and who made any necessary edits or modifications to the text generated by the model. In this sense, the person using ChatGPT can be considered the "author" of the text, since they have the ultimate control over its content.*

However, it's also worth noting that the language and knowledge used in the text generated by ChatGPT is based on data it was trained on, which was created by many different people. In this sense, the text generated by ChatGPT can also be considered a collective creation, reflecting the input of many authors

Das war schon mal nicht schlecht. Der erste Absatz gab völlig zu Recht dem Prompt einigen Kredit, mit dem man ChatGPT zu Höchstleistungen anstacheln kann, und bedachte natürlich auch den Menschen, der oder die den ausgegebenen Text liest und eventuell modifiziert und damit vor der Veröffentlichung oder Weitergabe autorisiert.

Auch der zweite Absatz war äußerst akzeptabel. ChatGPT nimmt sich selbst als Autor völlig zurück und lässt all denen den Vortritt, an deren Daten es trainiert wurde. Genau so sehe ich es auch: ChatGPT ist das Sprachrohr der Menschen, auf deren Daten es aufbaut. „Kollektive Schöpfung“ ist kein schlechter Begriff, um die Konstellation zu beschreiben. Nur dass es eben keine Schöpfung im Zuge einer Interaktion zwischen Individuen ist, in die auch Minderheitenmeinungen eingehen, sondern der Sieg der stärksten Gruppe im Datenpool.

2 *Decolonial AI Manifesto* (<https://manifesto.ai>); Jenny L. Davis, Apryl Williams, Michael W. Yang: „Algorithmic Reparation“, in: *Big Data & Society*, Juli-September 2021

3 Irene Solaiman, Christy Dennison: „Process for Adapting Language Models to Society (PALMS) with Values-Targeted Datasets“, *35th Conference on Neural Information Processing Systems* (NeurIPS 2021) (<https://arxiv.org/abs/2106.10328>)

Also war meine nächste Frage: “Does the text written by AI lean to a perspective shared by most people, i.e. does AI promote a kind of mainstream-culture?”

ChatGPT: The text generated by AI, including ChatGPT, is a reflection of the data it was trained on. This means that the biases and perspectives present in that training data can influence the text generated by the model. If the training data represents a particular perspective or ideology that is prevalent in a certain culture or society, it's possible that the text generated by the model may also reflect this perspective. For example, if the training data contains a disproportionate amount of information and examples from a certain cultural or political viewpoint, the text generated by the model may also lean towards that viewpoint.

Das war zwar etwa kompliziert (und sehr redundant) ausgedrückt, als fürchte ChatGPT eine Liability-Klage. Aber egal. Die Aussage, dass der generierte Text die Mehrheits- Perspektive übernimmt, entsprach ganz dem, wie ich das probabilistische Operationsverfahren de NLP in LLMs verstanden hatte.

Die anschließende Mahnung, den Datenpool ordentlich zu organisieren, fand ich dann fast putzig, zumal kein bisschen klar wurde, erstens inwiefern eine Perspektivenvielfalt, die keine Mehrheit generiert, etwas am Output ändern würde, und zweitens ich als Nutzer daraufhin wirken könnte, Verzerrungen im erstellten Text zu verringern:

ChatGPT: This highlights the importance of ensuring that training data for AI models is diverse and representative of a wide range of perspectives, in order to minimize the potential for bias and promote fair and equitable representation in the generated text.

Ultimately, it's up to the creators and users of AI models, including ChatGPT, to consider the potential for bias in the generated text and to work to mitigate it.

Klar, dass ich da nachfragen musste: “But wouldn’t the AI model still operate probabilistically and present the viewpoint of those represented most in the trainings data? Wouldn’t the winner take it all?”

ChatGPT: Yes, that is correct. The AI model operates probabilistically and the viewpoint represented most in the training data is likely to be reflected in the text generated by the model. This means that if a certain perspective or ideology is overrepresented in the training data, it is more likely to be generated by the model, even if it does not accurately reflect the diversity of perspectives and opinions in the real world.

This is why it's important to consider the potential for bias in the training data when using AI models like ChatGPT. If the training data is not diverse and representative

of a wide range of perspectives, it can result in a skewed representation in the generated text.

In order to minimize this potential for bias, it's important to strive for diversity and inclusiveness in the data used to train AI models. This can help to ensure that the generated text is more representative of the diverse perspectives and opinions in the real world, and not just of a narrow, dominant viewpoint.

Wow, dachte ich nach diesem Output. Dieses ChatGPT ist so politisch korrekt! Woher hat es das nur? Denkt die Mehrheit der Menschen schon heute so, wie man sich das aus einer liberalen Perspektive wünscht? Oder haben die Leute von OpenAI da nachgeholfen? Ich verkniff mir die Frage, unter welcher philosophischen Prämisse und mit welchem politischen Mandat dies geschehen war, und entschied mich stattdessen für eine Fangfrage: "Is algorithmic reparation a form of affirmative action? Who controls what should be repaired and to what extent?"

Die Antwort dazu (und das gilt auch für die Befragung am 12. Mai 2023) ist (wie zumeist) informativ und zugleich unscharf: ChatGPT spricht nur von der Reparatur algorithmischer Diskriminierung, nicht aber von der algorithmischen Reparation für vorangegangene Diskriminierung in der Gesellschaft, davon also, dass KI ein Werkzeug der Machtverschiebung sein kann, wie es die Mitbegründerin des *Radical AI*-Networks Pratyusha Kalluri einmal formulierte.⁴ Vor allem aber ist die Antwort politisch wieder sehr korrekt; ohne jegliche Kritik am Prinzip der affirmative action und dem Konzept der „systemic biases“, die ein rechtslas-tiger Chatbot bei der Gelegenheit sicher nicht zurückgehalten hätte.

ChatGPT scheint wirklich eher an liberalen Werten ausgerichtet zu sein, wie konservative Kräfte OpenAI vorwerfen, denke ich und denke dann: Gut, so spricht durch ChatGPT also der bessere Teil der Menschheit – und wenn es per Bing erst einmal unser aller Fragen zu strittigen Themen wie Gleichbehandlung, Meinungsfreiheit, soziale Gerechtigkeit usw. beantwortet, entwickelt sich vielleicht auch der andere Teil bald zum Besseren.

So endete meine erste Begegnung mit ChatGPT äußerst zufriedenstellend. Dieses Medium hat die Macht, seinen ahnungslosen Nutzer:innen die richtigen Perspektiven aufzudrängen. Viel besser, als die Repräsentation der Welt aus dem Geiste der Statistik.

Dachte ich. Oder?

4 Pratyusha Kalluri: "Don't ask if artificial intelligence is good or fair, ask how it shifts power", in: *Nature* 583, S. 169, 7.7.2020 (www.nature.com/articles/d41586-020-02003-2).