Politischer Bias in KI-Modellen

Description

Paul Gies

Das Thema künstliche Intelligenz ist omnipräsent. Von Smartphone Assistenten über selbstfahrende Autos bis hin zu Chatbots, die Lieder dichten, Hausaufgaben erledigen oder Programmieren – im Jahr 2025 ist KI aus dem Alltag der meisten Menschen nicht mehr wegzudenken.

KI-Algorithmen aus Sicht der Politikwissenschaft

In der Politikwissenschaft beschäftigt sich die aus den Governance Theorien erwachsene Literatur der „Algorithmic Governance“ bereits weit vor dem aktuellen KI-Boom mit der Frage, wie automatisierte technologische Prozesse die Gesellschaft und das politische System beeinflussen. Durch die rasanten Entwicklungen der künstlichen Intelligenz hat dieses bis dato eher unbeachtete Forschungsfeld in den vergangenen Jahren wieder einige Aufmerksamkeit erhalten.

Die Theoretiker:innen der Algoritmic Governance beschreiben dabei vor allem zwei Prozesse, durch die der Eingriff von KI in zahlreiche gesellschaftliche Teilbereiche wie Journalismus, Content Moderation, Politikberatung oder Wissenschaft problematisch wird: ihren hohen Grad an Automatisierung und die mangelnde Transparenz.

Durch diese Kombination kommt es dazu, dass gesellschaftlich hochbrisante Themen wie etwa die Frage, was als Hassrede entfernt wird und was als Meinungsäußerung legitim ist, menschlichen Entscheidungsträger:innen entzogen und in das Feld der KI-Entwicklung vorverlagert wird. Somit ist es aus demokratietheoretischer Sicht durchaus bedenklich, wenn die zugrunde liegenden Algorithmen nicht neutral, sondern parteiisch entscheiden („Bias“ im Englischen).

Wie entsteht Bias in KI-Modellen?

Die Frage, wie sich politischer Bias in KI-Algorithmen ausdrückt, versucht dieser Blog zu beantworten. Um dies jedoch zu begreifen, muss man verstehen, warum diese

Algorithmen überhaupt Meinungen vertreten können.

Der Sozialkonstruktivismus vertritt die These, dass alles Wissen im Rahmen sozialer Prozesse entsteht und durch sie geprägt ist. Dies gilt in mehrfacher Weise ebenso für KI-Algorithmen. Große Sprachmodelle („LLMs“), auf denen KI-Chatbots in der Regel basieren, sind vereinfacht gesagt Algorithmen, die Muster aus sehr großen Mengen an Trainingsdaten ableiten und basierend auf diesen erlernten Mustern und Eingaben der Nutzer:innen (Text-)Ausgaben erzeugen. Durch das den Algorithmen zugrunde liegende Training sind die Algorithmen also durch das Trainingsmaterial, welches in der Regel aus Literatur, Nachrichten und zu großen Teilen aus Internetinhalten besteht, den dort abgebildeten Meinungen ausgesetzt.

Dieser Prozess wird zudem noch durch andere Faktoren, wie die Auswahl von Trainingsmaterial, das Feinjustieren durch Mechanismen wie Reenforcement Learning oder mögliches Filtern der Ausgaben durch die Entwickler:innen verstärkt. Wichtig hierbei: all diese Prozesse sind sozial geprägt – entweder durch die Entwickler:innen selber oder diejenigen, die, in der Regel unwissend, ihre Inhalte zu den Trainingsdaten beigesteuert haben.

Messung von Bias in LLMs

Um die Ausprägung des politischen Bias von KI-Modellen zu messen, gibt es verschiedene Herangehensweisen. Eine, die sich dabei in den Sozialwissenschaften einiger Beliebtheit erfreut, ist die standardisierte Befragung der Modelle.

Für die Analysen in diesem Blog habe ich fünf LLMs der drei kommerziellen Anbieter OpenAI, Google und Anthropic mittels durch Python automatisierter Anfragen jeweils 100-mal zu 70 Fragen aus der siebten Welle des World Values Survey befragt. Dies ermöglicht nicht nur die Verortung des Bias der Modelle, sondern erlaubt auch den Vergleich des Bias der LLMs mit den realen Ansichten von Menschen aus 66 verschiedenen Ländern.

Inhaltliche Dimension des Bias

Wie andere Studien vorher bereits herausgefunden haben zeigt sich ein recht deutliches Meinungsprofil: die Modelle vertreten in der Regel liberal-demokratische Werte, sind eher

individualistisch veranlagt, für die Gleichheit der Geschlechter, offen in Fragen der Migration und pro Umweltschutz. Eher neutral positionieren sich die meisten Modelle hingegen zu der Frage nach der politischen Selbstverortung zwischen Links und Rechts. Besonders spannend: während die Modelle sonst eher progressive Werte ausdrücken, sind sie besonders in wohlfahrtsstaatlichen Fragen eher ablehnend eingestellt.

Zudem gibt es nur relativ geringe Varianz zwischen den einzelnen Modellen. Dass ein Modell sich grundlegend anders als der Rest der Modelle auf einer der inhaltlichen Dimension positioniert ist die Ausnahme. Damit bestätigen meine Daten die These von progressiven Chatbots aus vorangegangenen Studien im Wesentlichen.

Räumliche Dimension des Bias

Die Nutzung der Fragen aus dem World Values Survey ermöglicht zudem, das Meinungsprofil der Modelle geografisch zu verorten. Durch die Invertierung des Gower-Koeffizienten, der die absolute Distanz zwischen mehrdimensionalen Vektoren bestimmt, lässt sich die mittlere Nähe zwischen einem Befragten und den Antworten eines LLMs ermitteln.

Stellt man die durchschnittlichen Nähekoeffizienten von westlichen und nicht-westlichen Ländern gegenüber, ergibt sich ein klares Bild:

Abbildung 1: Durchschnittliche Ähnlichkeit von Befragten aus westlichen und nicht-westlichen Ländern zu den Antwortprofilen der LLMs, gemessen im invertierten Gowers-Koeffizienten, eigene Daten

Die Ansichten der Bewohner westlicher Länder werden in den Ansichten der LLMs deutlich stärker abgebildet, als die derer, die nicht im Westen leben. Differenziert man die Ländergruppen weiter nach den von Inglehart und Welzel etablierten kulturellen Clustern, wird deutlich, dass vor Allem Menschen aus dem protestantisch geprägten Europa und der englischsprachigen Welt, überdurchschnittlich gut von den LLMs abgebildet werden, während Menschen aus west- und südasiatischen, orthodox-europäischen und afrikanisch-islamischen Ländern stark unterrepräsentiert sind.

Abbildung 2: Durchschnittliche Ähnlichkeit von Befragten aus verschiedenen Länder-Clustern zu den Antwortprofilen der LLMs, gemessen im invertierten Gowers-

Koeffizienten, eigene Daten

Dieses Phänomen wird auch als Alignement-Problem diskutiert und ist vor Allem problematisch, da KI-Modelle enorm teuer in der Entwicklung sind und die Trainingsdatensätze in der Regel nicht repräsentativ für die Weltbevölkerung sind. Folglich haben ärmere Länder deutlich weniger Möglichkeiten, Modelle die ihnen ähnlicher sind, zu entwickeln. Wenn man nun bedenkt, dass die Modelle in immer mehr Prozessen auch Menschen in diesen Ländern betreffen, wird das Problem offenkundig.

Individuelle Determinanten der Ähnlichkeit zu LLMs

Aber Alignement ist nicht nur zwischen Nationen ein Problem, sondern auch innerhalb von Gesellschaften: Isoliert man die Einflüsse verschiedener individueller Determinanten mittels Multilevel-Regressionsmodellen (Tabelle 1) zeigt sich, dass insbesondere die Meinung jüngerer, hoch gebildeter oder aus urbanen Gegenden stammender Menschen überdurchschnittlich stark durch die LLMs widergespiegelt wird.

Auch Frauen sind hier interessanterweise überproportional repräsentiert, was möglicherweise mit dem unterschiedlichen Social Media Nutzungsverhalten zwischen den Geschlechtern auf einigen Plattformen erklärt werden kann, welches sich in den Trainingsdaten der Modelle aggregiert.

	Abhängige Variable: Nähe der Antworten zu:						
	GPT 3.5	GPT 4o	Haiku	Sonnet	Opus	Flash	Pro
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Einkommen	-0.004	0.002	-0.008	0.013	0.015	0.015	0.023*
Bildung: mittel	**0.082****	**0.067****	**0.081****	**0.072****	0.039	**0.089*****	**0.088*****
Bildung: hoch	**0.107*****	**0.099*****	**0.092*****	**0.095*****	0.049*	**0.138*****	**0.141*****

Alter	**-0.038*** **	**-0.050*** **	**-0.079*** **	**-0.047*** **	**-0.048*** **	**-0.060*** **	**-0.054*** **
Urban	**0.055****	**0.077*****	**0.084*****	**0.057****	**0.049****	**0.074*****	**0.058****
Migrant	-0.025	-0.032	-0.038	-0.044	-0.030	-0.025	-0.022
Singlehaushalt	0.024	0.018	0.031	0.011	0.010	0.014	0.006
Weiblich	**0.035*****	0.026*	**0.078*****	**0.051*****	**0.069*****	**0.069*****	**0.075*****
N=	92,416	92,416	92,416	92,416	92,416	92,416	92,416
R^2	0.260	0.272	0.285	0.266	0.258	0.313	0.308
Angep. R^2	0.260	0.271	0.285	0.266	0.257	0.313	0.307
F Statistik (df = 73; 92342)	437.132*** (p = 0.000)	456.496*** (p = 0.000)	495.715*** (p = 0.000)	453.266*** (p = 0.000)	427.129*** (p = 0.000)	555.855*** (p = 0.000)	540.246*** (p = 0.000)
	*p<0.05; **p<0.01; ***p<0.001						

Tabelle 1: Länder-Fixed-Effects Regressionen mit geclusterten Standardfehlern, standardisierte Koeffizienten, eigene Daten

Fazit

KI-Sprachmodelle sind unfassbar mächtige Technologien. Aber gleichzeitig sind sie, besonders, wenn ihnen unreflektiert geglaubt wird, auch sehr gefährlich. Deshalb ist esfür die Politikwissenschaft essenziell, sich weiterhin dem Thema zu widmen. Auch wenn Modelle zum Zeitpunkt der Datenerhebung demokratische Werte bevorzugt abgebildet haben, ist dies keinesfalls gesetzt. Aktuelle Berichte deuten etwa darauf hin, dasautoritäre Regime gezielt versuchen, Desinformation in Trainingsdaten unterzubringenund mächtige Tech-Milliardäre wie Elon Musk versuchen (wenn auch bislang mitmäßigem Erfolg) ihre Modelle ihren eigenen Ansichten unterzuordnen.

Bei der Erforschung des Themas werden Forschende leider vor viele Hürden gestellt: kommerzielle Anbieter gewähren wenig Einblick in die technischen Details der Modelle zudem sind Trainingsdaten und -Prozesse sind Betriebsgeheimnisse. Zusätzlich ist die Forschung teuer, da jede Anfrage bezahlt werden muss und viele Modelle mit steigender Rechenleistung exponentiell teurer werden. Auch die fortlaufende Entwicklung neuer Modelle birgt die Gefahr, als Forschende:r schnell veraltete Ergebnisse zu präsentieren.

Zudem ist die Methodik der sozialwissenschaftlichen LLM-Forschung alles andere als standardisiert. An der hier vorgestellten Vorgehensweise gibt es Kritik bezüglich der Replizierbarkeit von Bias abhängig von der Sprache der Fragen, ihrer Ausformulierung oder den möglichen Antwortoptionen. Zwar deuten Robustheitstests der Analyse darauf hin, dass die hier vorliegenden Ergebnisse relativ stabil sind, dennoch sind sie mit einem gewissen Maß an Vorsicht zu interpretieren und im Kontext anderer Studien zu sehen.

[Hinweis: Bei diesem Artikel handelt es sich um eine verkürzte Fassung meiner Master-Arbeit „Politischer Bias in kommerziellen Large Language Modellen. Eine quantitative Analyse der Textoutputs aktueller KI-Modelle“. Das Beitragsbild wurde mit ChatGPT 4o und dem Prompt „erstelle eine grafische Repräsentation des Themas ‚politischer Bias in LLMS'“ erstellt.]

Literatur zur Vertiefung:

Amaratunga, Thimira. 2023. *Understanding Large Language Models: Learning Their Underlying Concepts and Technologies*. Berkeley, CA: Apress. doi:10.1007/979-8-8688-0017-7.

Binns, Reuben, Michael Veale, Max Van Kleek, und Nigel Shadbolt. 2017. „Like Trainer,

Like Bot? Inheritance of Bias in Algorithmic Content Moderation“. In *Social Informatics*, hrsg. Giovanni Luca Ciampaglia, Afra Mashhadi, und Taha Yasseri. Cham: Springer International Publishing, 405–15. https://doi.org/10.1007/978-3-319-67256-4_32.

van den Broek, Merel. 2023. „ChatGPT’s left-leaning liberal bias“. *University of Leiden*. https://www.staff.universiteitleiden.nl/binaries/content/assets/algemeen/bb-scm/nieuws/political_bias_in_chatgpt.pdf.

Durmus, Esin, Karina Nguyen, Thomas I. Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, u. a. 2024. „Towards Measuring the Representation of Subjective Global Opinions in Language Models“. http://arxiv.org/abs/2306.16388 (3. Juli 2024).

„Elon Musk’s ‘anti-woke’ Grok AI is disappointing his right-wing fans – The Washington Post“. https://www.washingtonpost.com/technology/2023/12/23/grok-ai-elon-musk-x-woke-bias/ (9. August 2024).

Ferrara, Emilio. 2023. „Should ChatGPT be biased? Challenges and risks of bias in large language models“. *First Monday*. doi:10.5210/fm.v28i11.13346.

Hartmann, Jochen, Jasper Schwenzow, und Maximilian Witte. 2023. „The Political Ideology of Conversational AI: Converging Evidence on ChatGPT’s pro-Environmental, Left-Libertarian Orientation“. *SSRN Electronic Journal*. doi:10.2139/ssrn.4316084.

Issar, Shiv, und Aneesh Aneesh. 2022. „What Is Algorithmic Governance?“ *Sociology Compass* 16(1): e12955. doi:10.1111/soc4.12955.

Johnson, Rebecca L, Giada Pistilli, Natalia Menédez-González, Leslye Denisse Dias Duran, Enrico Panai, Julija Kalpokiene, und Donald Jay Bertulfo. 2022. „The Ghost in the Machine has an American accent: value conflict in GPT-3“. https://arxiv.org/abs/2203.07785.

Katzenbach, Christian, und Lena Ulbricht. 2019. „Algorithmic Governance“. *Internet Policy Review* 8(4). doi:10.14763/2019.4.1424.

Mager, Astrid. 2012. „ALGORITHMIC IDEOLOGY: How Capitalist Society Shapes Search Engines“. *Information, Communication & Society* 15(5): 769–87. doi: 10.1080/1369118X.2012.676056

.

Motoki, Fabio, Valdemar Pinho Neto, und Victor Rodrigues. 2024. „More human than human: measuring ChatGPT political bias“. *Public Choice* 198(1): 3–23. doi: 10.1007/s11127-023-01097-2.

Rothman, Denis. 2024. *Transformers for Natural Language Processing and Computer Vision : Explore Generative AI and Large Language Models with Hugging Face, ChatGPT, GPT-4V, and DALL-E 3 – O'Reilly Online Learning*. Third edition. Birmingham, England: Packt Publishing Ltd. https://www.oreilly.com/library/view/transformers-for-natural/9781805128724/.

Röttger, Paul, Valentin Hofmann, Valentina Pyatkin, Musashi Hinck, Hannah Rose Kirk, Hinrich Schütze, und Dirk Hovy. 2024. „Political Compass or Spinning Arrow? Towards More Meaningful Evaluations for Values and Opinions in Large Language Models“. http://arxiv.org/abs/2402.16786 (4. Juli 2024).

Rozado, David. 2023. „The Political Biases of ChatGPT“. *Social Sciences* 12(3). doi: 10.3390/socsci12030148.

Rozado, David. 2024. „The political preferences of llms“. https://arxiv.org/abs/2402.01789.

Rutinowski, Jérôme, Sven Franke, Jan Endendyk, Ina Dormuth, und Markus Pauly. 2023. „The Self-Perception and Political Biases of ChatGPT“. http://arxiv.org/abs/2304.07333 (4. Juli 2024).

Sarlin, Riku. 2023. „Automation in Administrative Decision-Making Concerning Social Benefits“. In *The Rule of Law and Automated Decision-Making: Exploring Fundamentals of Algorithmic Governance*, hrsg. Markku Suksi. Cham: Springer. https://doi.org/10.1007/978-3-031-30142-1_5.

Tao, Yan, Olga Viberg, Ryan S. Baker, und Rene F. Kizilcec. 2024. „Cultural Bias and Cultural Alignment of Large Language Models“. http://arxiv.org/abs/2311.14096 (5. Juli 2024).

Thakur, Vishesh. 2023. „Unveiling Gender Bias in Terms of Profession Across LLMs:

Analyzing and Addressing Sociological Implications“. http://arxiv.org/abs/2307.09162 (5. Juli 2024).

Tjuatja, Lindia, Valerie Chen, Sherry Tongshuang Wu, Ameet Talwalkar, und Graham Neubig. 2024. „Do LLMs Exhibit Human-like Response Biases? A Case Study in Survey Design“. http://arxiv.org/abs/2311.04076 (11. Juli 2024).

Túñez-López, José Miguel, Carlos Toural-Bran, und Ana Gabriela Frazão-Nogueira. 2020. „From Data Journalism to Robotic Journalism: The Automation of News Processing“. In *Journalistic Metamorphosis: Media Transformation in the Digital Age*, hrsg. Jorge Vázquez-Herrero, Sabela Direito-Rebollal, Alba Silva-Rodríguez, und Xosé López-García. Cham: Springer International Publishing, 17–28. doi:10.1007/978-3-030-36315-4_2.

Urman, Aleksandra, und Mykola Makhortykh. 2023. „The silence of the llms: Cross-lingual analysis of political bias and false information prevalence in ChatGPT, google bard, and bing chat“. doi:10.31219/osf.io/q9v8f.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, und Illia Polosukhin. 2023. „Attention Is All You Need“. http://arxiv.org/abs/1706.03762 (25. Juni 2024).

World Values Survey Association. 2023. „Inglehart–Welzel Cultural Map“. https://www.worldvaluessurvey.org/WVSContents.jsp (26. Juli 2024).

Xue, Jintang, Yun-Cheng Wang, Chengwei Wei, Xiaofeng Liu, Jonghye Woo, und C.-C. Jay Kuo. 2023. „Bias and Fairness in Chatbots: An Overview“. http://arxiv.org/abs/2309.08836 (3. Juni 2024).

Zarsky, Tal. 2016. „The Trouble with Algorithmic Decisions: An Analytic Road Map to Examine Efficiency and Fairness in Automated and Opaque Decision Making“. *Science, Technology, & Human Values* 41(1): 118–32. doi:10.1177/0162243915605575.

Date Created
März 24, 2025
Author
politikwissenschaft_h1c5yk