Wissenschaftliche Integrität und der Umgang mit Plattformen generativer Künstlicher Intelligenz in der Politikwissenschaft

Achim Goerres, diese Version 10.4.2024

Definitionen

Generative Künstliche Intelligenz ist eine digitale Prozedur, die ein Output generiert (vor allem Text, Audio, Bilder, Video), zu dessen Erstellung gleichzeitig Regeln des Users und des Algorithmus unter Berücksichtigung einer großen Trainingsdatenmenge und einer relativ kleinen User-Datenmenge eingesetzt werden. Das, was der User eingibt, nennen wir **Input** oder **Prompt**. Das, was die KI produziert, nennen wir **Output**. Den Prozess der Bedienung nennen wir **Prompting**. Die Daten, mit denen die unterliegenden Algorithmen ausgebildet worden sind, nennt man **Trainingsdaten**.

Chat GPT und Google Bard sind generative Sprachmodelle. Sie erstellen als Output die wahrscheinlichsten Kombinationen von Wörtern gegeben die grammatikalischen Regeln der verlangten jeweiligen Sprache und den Kontext des Prompts. Zusätzlich werden Zufallsentscheidungen genutzt. Dabei gehen sie vereinfacht gesagt seriell Wort für Wort vor. Diese Modelle können vorhersagen, dass der Satz

„Did you hear my covert narcissism I disguise…"

wie folgt endet:

„…as altruism like some kind of congressman?".

Diese Modelle (wenn sie Trainingsdaten bis 2022 haben) können es vorhersagen, weil er einem der populärsten Lieder im Jahr 2022-24 entnommen ist („Anti-Hero" von Taylor Swift), dessen Text in unzähligen Variationen seit 2022 im Internet zu finden ist. Sie können aber niemals diesen originellen Satz erstproduzieren, der völlig unterschiedliche sprachliche Konzepte erstmalig zusammenbringt. Deswegen sind diese Sprachmodelle unheimlich stark darin, Standardwissen zu reproduzieren, aber nicht zu originellem Output fähig – ganz im Gegenteil: sie setzen die Wörter ein, die am wahrscheinlichsten sind – und das ist genau das Gegenteil von Originalität. Dabei reproduzieren Sie hauptsächlich das "Internetwissen", also das Wissen der Welt, das im Internet aufbereitet ist. Dieses Wissen ist in wenigen Sprachen konzentriert und hat bestimmte regionale Schwerpunkte und Biases.

Diese Sprachmodelle sind nicht wirklich kreativ, weswegen das Adjektiv generativ nur im Sinne von "erstellend" zu verstehen ist.

Grundsätzliche Prinzipien von Wissenschaft

- Präzision
- Transparenz
- Systematisches, regelgeleitetes Vorgehen
- Wiederholbarkeit
- Objektivität
- Beschreibung der Leistungen Dritter (Personen, Quellen, Prozesse)

Verletzungen der Prinzipien von Wissenschaft beim Einsatz Generativer KI

Generative KI-Prozesse sind nicht identisch wiederholbar. Durch jeden Einsatz von Generativen KI-Prozessen verändert sich der Prozess selbst. Die Algorithmen beinhalten Regeln zur Anpassung der Schritte, die sie anwenden. Damit kann dasselbe Input in einen Prozess zu unterschiedlichen Zeitpunkten zu unterschiedlichen Outputs führen und wird jedes Mal in einen zu diesem Zeitpunkt einzigartigen Prozess eingespeist.

Kommerzielle Generative KI-Plattformen (wie Chat GPT und nur die nutzen wir zurzeit in der Politikwissenschaft) verstoßen häufig (nicht immer) gegen das Prinzip der Transparenz, wobei die Entwicklung sehr schnell ist und solch eine Bewertung schnell hinfällig sein könnte. Die Unternehmen müssen nicht zu 100 % deklarieren, wie ihre Angebote entstehen, wobei der EU-AI-Act dies bald ändern wird. Man kann in so einem Fall als Wissenschaftler*in nicht „unter die Haube" gucken. Deswegen gibt es kommerzielle und nicht-kommerzielle Open-Source-Angebote generativer KI, in denen ein Maximum an Transparenz verfolgt wird.

De facto verfügen aber nur sehr wenige Menschen in der Politikwissenschaft über die informatische Expertise, die Prozesse solcher Open-Source-Plattformen nachzuvollziehen. Aber selbst solche Experten können für konkrete Inputeingaben nicht nachvollziehen, warum der Prozess ein bestimmtes Ergebnis produziert.

Viele Politikwissenschaftler*innen bedienen deswegen de facto eine Black Box beim Einsatz von Generativer KI, deren Prozesse sie sich nicht nachvollziehen und deswegen auch nicht nachvollziehbar und transparent beschreiben können.

Ein weiteres gravierendes Problem ist die nicht zu gewährleistende Anerkennung von Leistungen Dritter. Die aktuell zugänglichen Plattformen Generativer KI nutzen als Trainingsdaten die Werke unzähliger, lebendender und verstorbenener Autor*innen mit einer starken Fokussierung auf Englisch (und wenigen anderen Sprachen, darunter Deutsch) und der damit einhergehenden Vernachlässigung von Werken in anderen Sprachen. Über diesen Ansatz kann der individuelle Beitrag der ursprünglichen Urheber*innen und der Menschen, die diese Werke verfügbar gemacht haben, nicht mehr gekennzeichnet werden. Damit können die Leistungen Dritter nicht präzise und transparent gekennzeichnet werden.

Ethisch integrer Einsatz von Generativer KI in der Politikwissenschaft

Das Output von Generative Künstlicher Intelligenz selbst kann niemals wissenschaftlichen Standards entsprechen. Die Prinzipien Nachvollziehbarkeit, Anerkennung der Leistungen Dritter und Wiederholbarkeit sind verletzt.

Grundsätzlich müssen sich Politikwissenschaftler*innen bei der Nutzung von kommerziellen KI-Plattformen darüber informieren, ob ihr Input zum weiteren Training des KI-Modells genutzt wird (und sie damit den Firmen mit ihrer Input-Leistung kostenlos einen Vorteil verschaffen) und ob sie damit einverstanden sind. Aktuell erlauben manche kostenpflichtige Plattformen den eigenen Input davon auszunehmen. Ethisch ist es unbedenklich, wenn Wissenschaftler*innen sich zur Weitergabe ihres eigenen Inputs entschließen. Dabei dürfen sie keinen Input von Dritten verwenden, es sei denn diese Dritten haben der Nutzung in diesem Sinne ausdrücklich zugestimmt.

Generative Künstliche Intelligenz kann bei der Erarbeitung eines wissenschaftlichen Textes helfen. Wissenschaftler*innen können sich den inhaltlichen Wissensstand in einem neuen Gebiet damit erarbeiten. Sie müssen aber immer alles selbst noch einmal prüfen, weil die Sprachmodelle keine Wahrheit produzieren können und manchmal scheinbar etwas erfinden, weil sie nach Wahrscheinlichkeiten zusammensetzen. Mit dieser Vorgehensweise nutzen Wissenschaftler*innen zwar eine Black Box, die sie nicht verstehen und nicht beschreiben können, aber sie hinterfragen das Output aus diesem Prozess.

Generative Künstliche Intelligenz kann in der Erarbeitung einer empirischen Analyse unterstützen, indem sie beispielsweise erste Versionen von Code produziert, der dann kritisch hinterfragt und erweitert wird.

Generative Künstliche Intelligenz kann zudem die Qualitätssicherung in der Produktion von wissenschaftlichem Text und Syntax unterstützen. Wissenschaftler*innen können von ihnen vorproduzierte Texte oder Code mithilfe von KI überprüfen lassen. Bei solch einer Qualitätskontrolle von menschlich produziertem Material können die Modelle ihre größte Stärke ausspielen: denn häufig liegen Fehler in der Text- oder Codeproduktion darin, dass man bestimmte Regeln nicht eingehalten hat und damit eine unwahrscheinlichere Kombination gewählt hat.

Der Output von Generativer Künstlicher Intelligenz kann selbst Gegenstand einer kritischen Analyse der Politikwissenschaft sein.

Notwendige Dokumentation eines ethisch integren Einsatzes in der Politikwissenschaft

- Sie können solche Dienste nutzen, müssen dies aber kenntlich machen, d.h. hinweisen und eine entsprechende Referenz anführen – im Sinne transparenten und nachvollziehbaren wissenschaftlichen Arbeitens. Bereiten Sie einen digitalen Anhang vor, der Einzelheiten über die verwendete KI-Services, das Datum der Verwendung, wörtliche Kopien des Inputs und Outputs enthält. Der Anhang kann schnell mehrere Dutzend Seiten lang werden.
- Sie sind verpflichtet, die konkrete Eingabeaufforderung (Prompt), deren Zeitpunkt sowie die Ausgabe des Programms zu dokumentieren. Tun Sie dies am besten mit einem Screenshot. Wenn Sie eine feinere Prompting-Technik anwenden, bei der Sie in einen "Dialog" mit der Schnittstelle treten, dokumentieren Sie nur die Prompts von Ihrer Seite und die endgültige Ausgabe.
- Wenn Sie einen wörtlichen Text aus der Ausgabe der Schnittstelle verwenden, müssen Sie dies durch Anführungszeichen und in einer Fußnote zum wörtlichen Zitat oder einer In-Text-Zitation kenntlich machen.
- Wenn eine Passage in Ihrem Aufsatz von der KI-Ausgabe beeinflusst wurde, führen Sie dies im Appendix auf: z.B. „Zeilen 43-46: der Text wurde mit Deepl Write verbessert.“ (Die Eingabe in Deepl Write und die Ausgabe von Deepl Write werden dann zusätzlich in den digitalen Anhang aufgenommen.

Bei der Erarbeitung dieses Handouts habe ich Input eines institutsinternen Leitfadens (Version 25.4.2023) und von Paul Gies, Teresa Hummler, Philipp Kemper, Johanna Plenter, Henrik Schillinger, Oliver Schwarz, Paul Vierus, Conrad Ziller genutzt, denen ich herzlich danke.