# Die unsichtbare Verschiebung unserer Meinungen durch KI


KI-Systeme wie ChatGPT können Perspektiven verzerren - ein Risiko, das Europas Regulierung bislang nur teilweise erfasst

Große Sprachmodelle beeinflussen zunehmend, wie Menschen Informationen wahrnehmen und bewerten. Die Studie "Communication Bias in Large Language Models: A Regulatory Perspective", erschienen im Journal "Communications of the ACM", zeigt nun, dass diese Systeme gesellschaftliche und politische Perspektiven verzerren können und analysiert zugleich, warum die europäische KI-Regulierung dieses Risiko bislang nur teilweise erfasst.

Die systematische Bevorzugung, Verstärkung oder Ausblendung bestimmter sozialer, kultureller oder politischer Perspektiven in KI-generierten Antworten bezeichnen die Forscher als Kommunikationsbias. "Sprachmodelle liefern nicht nur Informationen. Sie strukturieren auch, welche Argumente sichtbar werden und welche Deutungen plausibel erscheinen. Diese oft subtilen Effekte sind für Nutzer*innen kaum sichtbar und können öffentliche Debatten und demokratische Meinungsbildung beeinflussen", sagt Stefan Schmid, Leiter des Fachgebiets "Internet Architecture and Management" an der TU Berlin und Projektleiter am Weizenbaum-Institut, der die Studie mit Adrian Kuenzler, Professor an der Rechtswissenschaftlichen Fakultät der University of Hong Kong, durchgeführt hat. Zusammen haben sie untersucht, wie solche Verzerrungen in großen Sprachmodellen entstehen und welche Herausforderungen sich daraus für Regulierung und Steuerung ergeben.

## Wie Kommunikationsbias entsteht

Die Autoren identifizierten mehrere Ursachen für kommunikative Verzerrungen in großen Sprachmodellen. Eine wichtige Rolle spielen Trainingsdaten, in denen bestimmte Perspektiven stärker vertreten sind als andere. Auch Entscheidungen beim Training und Feintuning der Modelle beeinflussen, welche Argumentationsmuster bevorzugt werden.

Hinzu kommt eine weitere Eigenschaft vieler Sprachmodelle: Sie neigen dazu, sich an Erwartungen oder Positionen von Nutzer*innen anzupassen ? ein Effekt, der in der Forschung als Sycophancy bezeichnet wird. Da diese Einflüsse häufig kontextabhängig und schwer messbar sind, lassen sich solche Verzerrungen bislang nur begrenzt systematisch erfassen.

## Interdisziplinäre Analyse von Technik, Recht und Markt

Für ihre Studie verbinden Schmid und Kuenzler technische, rechtliche und ordnungspolitische Perspektiven. Sie ordnen bestehende Forschung zu politischen Verzerrungen, Polarisierung und Sycophancy in Sprachmodellen ein und analysieren zentrale europäische Regelwerke: den AI Act, den Digital Services Act (DSA) und den Digital Markets Act (DMA).

Ihr Befund: Kommunikationsbias ist kein Randphänomen, sondern ein strukturelles Thema an der Schnittstelle von Technikdesign, Marktstruktur und Regulierung.

Warum die Regulierung bislang nur begrenzt greift

Nach Einschätzung der Autoren setzen die bestehenden europäischen Regelwerke zwar wichtige Standards für Transparenz, Sicherheit und Rechenschaft. Die kommunikative Wirkung großer Sprachmodelle wird jedoch bislang meist nur indirekt adressiert. Sie betonen zudem, dass große Sprachmodelle zunehmend als eigenständige Untergruppe von KI-Systemen zu betrachten seien. Gerade weil sie für viele Nutzer*innen die unmittelbare Schnittstelle zur KI bildeten, könnten sie nicht nur öffentliche Debatten, sondern auch persönliche Entscheidungen in Bereichen wie Gesundheit, Finanzen und Politik beeinflussen. Nach Auffassung der Forscher erfassen die bestehenden Regelwerke dieses Risiko bislang nur unzureichend, insbesondere soweit es um den Einfluss auf grundlegende Weltbilder, soziale Perspektiven und politische Entscheidungsprozesse geht.

Der AI Act konzentriert sich vor allem auf Pflichten entlang der Entwicklungs- und Bereitstellungskette von KI-Systemen. Der Digital Services Act greift insbesondere dort, wo bereits konkrete systemische Risiken oder problematische Inhalte sichtbar werden. Kommunikationsbias erscheint damit eher als Nebenfolge anderer Regulierungsmechanismen als eigenständiges Problem öffentlicher Kommunikation.

## Gefahr der Marktkonzentration

Hinzu kommt die zunehmende Marktkonzentration im Bereich großer Sprachmodelle. Wenn wenige Unternehmen zentrale Modelle, Schnittstellen und Datenzugänge kontrollieren, können sich bestimmte kommunikative Muster besonders stark durchsetzen. Aus Sicht der Autoren muss deshalb neben der Inhalts- und Sicherheitsregulierung auch die Wettbewerbsordnung stärker in die Steuerung von KI einbezogen werden.

"Wer Risiken durch Sprachmodelle wirksam begrenzen will, darf nicht nur auf Verbote oder Einzelfallmoderation setzen", sagt Adrian Kuenzler. "Es braucht ein Informationsökosystem, in dem Vielfalt, Nachvollziehbarkeit und Wettbewerb technisch, regulatorisch und institutionell mitgedacht werden."

## Ein umfassender Steuerungsansatz

Die Studie plädiert für einen breiteren Regulierungsansatz. Dazu gehören Vorgaben entlang der gesamten Entwicklungs- und Nutzungskette, überprüfbare Aufsichts- und Moderationsmechanismen, mehr Wettbewerb im Markt für KI-Systeme sowie eine technische Gestaltung, die Vielfalt und Transparenz gezielt stärkt.

Die Autoren verweisen unter anderem auf regelmäßige Audits, geeignete Benchmarks zur Untersuchung kommunikativer Verzerrungen, größere Transparenz über Trainingsdaten und Modellverhalten sowie wirksame Korrekturmöglichkeiten nach der Markteinführung. Ziel sei ein digitales Kommunikationsumfeld, das Verzerrungen begrenzt und eine vielfältige öffentliche Debatte

unterstützt.

Über die Studie
Adrian Kuenzler, Stefan Schmid: Communication Bias in Large Language Models: A Regulatory Perspective. Communications of the ACM. DOI: 10.1145/3769689

https://dl.acm.org/doi/10.1145/3769689