KI-Transparenz per Regulierung

Einfach erklärt am Beispiel der Grundlagenmodelle

Alexander Brink und Leonhard Henke

Mit der Verabschiedung des EU AI Acts ist ein zentrales Thema in den Fokus der öffentlichen Diskussion zu Künstlicher Intelligenz (KI) gerückt: die Transparenz von KI-Systemen und die Frage, an welchen Stellen in diesen komplexen Strukturen Transparenz notwendig ist, um die Grundrechte der Bürgerinnen und Bürger der EU zu wahren.

Diese Debatte wurde kürzlich auch auf dem Digitalgipfel in Jena beleuchtet, wo Vertreter*innen von Organisationen wie Algorithm Watch für eine stärkere Regulierung von Foundation Models plädierten. Diese Modelle bilden zunehmend die Grundlage für eine Vielzahl von KI-Systemen. Das Wirtschaftsministerium hingegen vertrat eine gegenteilige Position und sprach sich für eine Regulierung der Anwendung dieser Modelle und nicht der Modelle selbst aus. Auch wenn nun entschieden wurde, dass die Grundlagenmodelle in Zukunft reguliert werden bzw. „angemessenen und spezifischeren Anforderungen und Verpflichtungen unterliegen" – wie es im Gesetz heißt –, bleiben die zugrunde liegenden ethischen Herausforderungen und die Pro- und Contra-Argumente, die diese Positionen untermauern, von entscheidender Bedeutung für die Diskussion. Denn trotz der Verabschiedung des AI Act werden die Diskussionen bei der nun folgenden technischen Ausgestaltung und Umsetzung des Gesetzes, insbesondere im Hinblick auf das zentrale Thema der Transparenz, zweifellos weitergehen.

1. *Gegenstände der Diskussion: KI-Systeme, Foundation Models und General Purpose AI*

KI-Systeme sind laut EU-Definition Systeme, die mit einem gewissen Grad an Autonomie agieren, Daten und Eingaben verwenden, um bestimmte Ziele zu erreichen, und Ausgaben wie Inhalte oder Entscheidungen erzeugen, die ihre Umgebung beeinflussen. In diesem Kontext gewinnt Generative AI an Bedeutung, deren Hauptaufgabe die Generierung von Inhalten ist. Hierunter sind auch diejenigen komplexen Sprachmodelle einzuordnen, die als Large Language Models (LLMs) insbesondere in Form von ChatGPT seit Ende 2022 weitreichende öffentliche Aufmerksamkeit erhalten haben.

Solche Modelle können auch als Foundation Models (dt. Basismodelle) bezeichnet werden, wenn sie im Hinblick auf Allgemeinheit und Vielseitigkeit der Ergebnisse optimiert wurden. Sie werden auf Grundlage eines breiten Spektrums von Datenquellen und großen Datenmengen trainiert und können zur Entwicklung von KI-Systemen mit spezifischer Zweckbestimmung oder von sogenannten General Purpose AI (GPAI)-Systeme eingesetzt werden. GPAI-Systeme können in einem breiten Spektrum von Anwendungen eingesetzt und an diese angepasst werden, z. B. zur Bild- und Spracherkennung, Audio- und Videogenerierung oder zur Beantwortung von Fragen. Die Hauptmerkmale von Foundation Models und GPAI – ihr großer Umfang, ihre Undurchsichtigkeit und ihr Potenzial, unerwartete Fähigkeiten zu entwickeln, die über die von ihren Entwickler*innen beabsichtigten hinausgehen – werfen eine Reihe von Fragen auf.

2. *Eine Frage der Ethik*

Ethik ist die Reflexionstheorie von Moral. Sie reflektiert über die Qualität kollektiver Wertvorstellungen von Menschen über spezifische Situationen. Das Spektrum reicht von Immanuel Kants deontologischem Ansatz bis zu den gesellschaftlichen Hintergrundvorstellungen eines John Rawls. Die 17 Ziele nachhaltiger Entwicklung sind gegenwärtig der bekannteste globale Normenkatalog, auf den sich alle 193 Mitgliedsstaaten der Vereinten Nationen weltweit verständigt haben. Ethik befasst sich aber auch mit grundlegenden Prinzipien wie Gerechtigkeit, Gleichheit, Respekt und Fairness – große Themen, die zunehmend in den digitalen Raum und auf digitale Entscheidungssituationen verlagert werden. Die ethischen Herausforderungen im Zusammenhang mit KI-Systemen sind vielfältig und eines der drängends-

ten Fragestellungen im Rahmen der Digitalverantwortung („corporate digital responsibility"). Im Zentrum der Debatte stehen die KI-Transparenz durch Regulierung, insbesondere im Zusammenhang mit Foundation Models.

3. Ethische Herausforderungen von Foundation Models

Ethik ist eine Reflexionstheorie, deren Stärke in ihrem ständigen Abwägungs- und Kompromissprozess liegt. In der Öffentlichkeit gibt es hingegen häufig fokussierte Standpunkte, die einen einzelnen Aspekt besonders stark hervorheben und dabei – leider – größere Zusammenhänge, Nebenwirkungen, Langfrist- oder Reboundeffekte etc. außer Acht lassen. Wir wollen im Folgenden die einschlägigsten ethischen Herausforderungen darstellen.

1. Transparenz und Verantwortung: Die Komplexität von Foundation Models kann zu mangelnder Transparenz führen. Wenn die Funktionsweise des Modells nicht ausreichend verstanden wird, besteht das Risiko, dass Entscheidungen nicht nachvollziehbar sind, insbesondere in erst darauf aufbauenden KI-Systemen. Dies wirft Fragen hinsichtlich der Rechenschaftspflicht (ex post) und ethischer Verantwortung (ex ante) auf, besonders in sensiblen Bereichen wie Bildung, Beschäftigung, kritischer Infrastruktur, öffentlicher Dienst, Strafverfolgung, Grenzkontrolle und Justizwesen.
2. Vorurteile und Diskriminierung: Wenn Foundation Models auf großen und möglicherweise ungleich verteilten Datensätzen trainiert werden, können sie erlernte Vorurteile und Diskriminierung im KI-System reproduzieren, in dem sie eingesetzt werden. Diese Herausforderung spitzt sich zu, wenn den Ergebnissen durch die Fähigkeiten des Systems wissenschaftliche Glaubwürdigkeit verliehen wird.
3. Datenschutz und geistiges Eigentum: Der Einsatz von umfangreichen, oft persönlichen Daten und urheberrechtlich geschütztem Material in Foundation Models birgt Risiken für Datenschutz, Sicherheit und den Schutz geistigen Eigentums. Die breite Anwendung dieser Modelle in verschiedensten Branchen und Bereichen intensiviert diese Problematik.
4. Missbrauch und Manipulation: Foundation Models könnten, beabsichtigt oder unbeabsichtigt, potenziell für unethische Zwecke missbraucht werden. Dies reicht von der Verbreitung von Fehlinformationen bis hin zur Erstellung gefälschter Inhalte. Darauf aufbauende KI-Systeme könnte für Manipulationen genutzt werden, die die Integrität von Informationen gefährden.

5. Umweltauswirkungen: Das Training von großen Foundation Models erfordert erhebliche Rechenressourcen, was zu einem hohen Energieverbrauch führt. Dies kann Umweltauswirkungen haben und stellt Fragen zur Nachhaltigkeit und Verantwortung in Bezug auf den Ressourcenverbrauch von KI-Entwicklungen. Damit wird das Nachhaltigkeitsziel 13 „Maßnahmen zum Klimaschutz", der Leitstern unter den Nachhaltigkeitszielen, potenziell negativ beeinflusst.
6. Wettbewerbsverzerrung: Ethische Bedenken können auch im Wettbewerbskontext auftreten. Unternehmen, die Zugang zu fortschrittlichen Foundation Models haben oder auf darauf basierende Anwendungen, können ein erhebliches Effizienzpotenzial heben und Wettbewerbsvorteile erlangen. Kleinere Unternehmen oder Organisationen könnten Schwierigkeiten haben, mit den technologischen Entwicklungen Schritt zu halten.

4. Der Foundation Model Transparency Index

Der „Foundation Model Transparency Index" des Stanford Institute for Human-Centered Artificial Intelligence (HAI) der Universität Stanford bezieht sich auf die Transparenz von Foundation Models, also der ersten Herausforderung. Forscher*innen haben die Transparenz von Modellen untersucht, die von zehn führenden Unternehmen und Organisationen als Grundlage für verschiedene KI-Systeme verwendet werden. Der Foundation Model Transparency Index legt 100 Indikatoren fest, die die Transparenz für Basismodelle umfassend kodifizieren. Transparenz in diesem Zusammenhang bezieht sich darauf, wie gut die Funktionsweise dieser Modelle verstanden und nachvollzogen werden kann, insbesondere in Bezug auf Entscheidungen, die sie treffen. Die Ergebnisse sind erstaunlich: es gibt – so die Forscher*innen – gegenwärtig ein beträchtliches Transparenzdefizit. Die höchste erreichte Werte ist 54/100, der Durchschnitt liegt bei 37/100. Damit sind auch weitere Aspekte betroffen wie die Gefahr der Zunahme von Diskriminierungen oder der mangelnde Datenschutz.

Es ist wichtig, diese Herausforderungen zu adressieren, um sicherzustellen, dass KI-Technologien auf ethisch verantwortungsbewusste Weise entwickelt, implementiert und genutzt werden. Die Diskussion über ethische Richtlinien und Standards in der KI-Forschung und -Anwendung ist daher von entscheidender Bedeutung. Kürzlich wurde ein offener Brief von führenden KI-Forscher*innen veröffentlicht, der einige der zuvor skizzierten CONTRA-Argumente aufgreift (vgl. Zeit 2023). Die Forscher*innen setzen sich für harte Regulierungen zur Sicherstellung der

KI-Transparenz und lehnen eine eher weiche freiwillige Selbstverpflichtung ab.

Damit widersprachen sie der Bundesregierung schon vor der Einigung von Rat und Parlament zum AI Act und forderten diese auf, ihre Position zum AI Act zu revidieren. Bundesregierungsvertreter*innen hingegen betonten kürzlich, dass Regulierung am wirksamsten auf der Anwendungsebene statt auf der Modellebene erfolgen sollte. Foundation Models sollten – so die Autor*innen des offenen Briefes – in den AI-Act integriert werden.

5. Die TOP Pro-Argumente für mehr KI-Transparenz durch Regulierung

1. Verbrauchervertrauen stärken: Eine transparente KI-Entwicklung ermöglicht es Verbraucher*innen und Nutzer*innen, das Verhalten von KI-Systemen besser zu verstehen. Dies trägt zur Stärkung des Vertrauens in KI-Anwendungen bei. Vertrauen ist wesentlich für die Akzeptanz einer neuen Technik durch Verbraucher*innen.
2. Ethik und Verantwortlichkeit fördern: Transparenz in KI-Modellen hilft dabei, ethische Standards und Verantwortlichkeiten zu gewährleisten. Entwickler*innen können nachvollziehbar darlegen, wie Entscheidungen getroffen werden (ex post), was wichtig ist, um mögliche Vorurteile oder diskriminierende Elemente in den Modellen zu identifizieren und zu korrigieren. Außerdem können sie ethische Verantwortung übernehmen (ex ante). Eine umfassende Dokumentation der Funktionsweise und der trainierten Modelle kann die Transparenz verbessern. Tools und Visualisierungen können entwickelt werden, um Benutzer*innen Einblicke in den Entscheidungsprozess des Modells zu geben.
3. Forschung und Entwicklung vorantreiben: Transparenz fördert den Austausch von Informationen und Erkenntnissen in der KI-Forschung. Entwickler*innen können von den Erfahrungen anderer lernen, was zu schnelleren Fortschritten und einer besseren Entwicklung von KI-Technologien führen kann. Damit wird ökonomischer Fortschritt gestärkt.
4. Schutz vor Missbrauch: Transparente KI-Systeme können dazu beitragen, Missbrauch und unethische Anwendungen zu verhindern. Wenn Entwickler*innen offener über ihre Modelle kommunizieren, wird es schwieriger, KI-Technologien für schädliche Zwecke zu nutzen.
5. Menschen mitnehmen: Digitale Teilhabe gilt als ein wesentliches Prinzip. Je mehr Menschen die Funktionsweise Künstlicher Intelligenz verstehen, umso eher gelingt die digitale Transformation.

6. Die TOP Contra-Argumente gegen mehr KI-Transparenz durch Regulierung

1. Schutz von Geschäftsgeheimnissen wird gefährdet: Einige Unternehmen argumentieren, dass die Offenlegung von KI-Modellen ihre Geschäftsgeheimnisse preisgeben könnte. Dies könnte dazu führen, dass Wettbewerber von den Entwicklungen profitieren, ohne den gleichen Aufwand betrieben zu haben. Vertraulichkeit und Verschwiegenheit stehen gegen Transparenz.
2. Die Komplexität der Modelle zu groß: Einige KI-Systeme sind äußerst komplex und schwer verständlich. Eine vollständige Transparenz könnte für die meisten Nutzer*innen wenig sinnvoll sein und möglicherweise mehr Verwirrung stiften als Aufklärung. Es kommt zu einem „information overload", der genau das Gegenteil bewirkt: die Nicht-Akzeptanz neuer Technologien.
3. Es entstehen Wettbewerbsnachteile: Unternehmen könnten zögern, ihre Modelle transparent zu machen, aus Angst, dass dies zu Wettbewerbsnachteilen führen könnte. Wenn andere Unternehmen leicht auf ihre Modelle zugreifen können, könnte dies ihre Marktposition beeinträchtigen. Der Marktmechanismus wird ausgehebelt.
4. Es entstehen gravierende Sicherheitsrisiken: Eine zu umfassende Transparenz könnte Sicherheitsrisiken schaffen, indem potenzielle Angreifer leichteren Zugang zu den Funktionsweisen der Modelle erhalten. Dies könnte zu Missbrauch und Manipulation führen.

Es ist wichtig, einen ausgewogenen Ansatz zu finden, der sowohl die Interessen der Entwickler*innen als auch die Bedürfnisse der Gesellschaft berücksichtigt. Nach dem der AI Act nun verabschiedet ist, beginnt die eigentliche Arbeit, ihn technisch auszugestalten und auch in der Praxis erfolgreich umzusetzen. Hier liegt noch beträchtlicher Diskussionsbedarf – die gesellschaftliche Debatte hat gerade erst begonnen.

7. Der AI Act der Europäischen Union

Nach einem dreitägigen Verhandlungsmarathon von Europaparlament, Ministerrat und Europäischer Kommission (Trilog) ist Europa der erste Kontinent, der einen Standard zu transparenter und verantwortungsvoller Gestaltung von KI-Nutzung setzt. Die EU-Kommission hatte das Gesetz im April 2021 vorgeschlagen.

Kernpunkte sind neben einer überarbeiteten Definition von KI die Adressierung extraterritorialer Aspekte und die Nichtberücksichtigung bestimmter Bereiche wie der nationalen Sicherheit sowie die teilweise Nichtberücksichtigung von Open-Source-Modellen. Risikobasiert unterscheidet das Gesetz zwischen KI mit minimalem, hohem und inakzeptablem Risiko sowie spezifischem Transparenzrisiko. Inakzeptabel ist der Einsatz von KI z. B. zum Aufbau von Sozialkreditsystemen. Hochriskante KI ist solche, die in sensiblen Bereichen wie medizinischen Geräten und im Zusammenhang mit kritischen Infrastrukturen eingesetzt wird. Für hochriskante KI-Systeme gelten zentrale Anforderungen wie umfassende Dokumentationspflichten, menschliche Aufsicht oder Anforderungen an die Cybersicherheit, um die Wahrung der Grundrechte zu gewährleisten. Foundation Models und General Purpose AI-Systeme müssen klare Transparenzstandards erfüllen, einschließlich der Offenlegung von Energieverbrauch und Trainingsdaten.

Mit dem AI Act hat die EU einen entscheidenden Schritt zur Gestaltung verantwortungsvoller und transparenter KI-Systeme getan. Er setzt Maßstäbe für die Balance zwischen Grundrechtsschutz und Innovationsförderung und sendet damit ein starkes Signal, dass die EU nach wie vor international eine wichtige Rolle im Umgang mit innovativer Technologie spielt. Er zeigt aber auch Lücken auf: Die Bestimmungen für Foundation Models müssen verschärft werden, die Haftungsfrage bleibt weitgehend offen und eine parallele Investitions- zur Regulierungsoffensive wird schmerzlich vermisst. Die EU muss nun pragmatische Regelungen anbieten, die mit der rasanten technologischen Entwicklung Schritt halten, um effektiv zu bleiben. Trotz der bestehenden Herausforderungen schafft der AI Act eine solide Grundlage, auf der Europa aufbauen kann, um ein führender Standort für KI-Innovationen zu werden und KI im öffentlichen Interesse zu steuern. Durch die Klärung der offenen Fragen könnte Europa nicht nur seine Wettbewerbsfähigkeit erhalten, sondern auch mit „KI made in Europe" eine führende Position in KI-Entwicklung und -Anwendung einnehmen und durch höchste Qualitäts- und Sicherheitsstandards international Investoren anziehen. Angesichts des zunehmenden globalen Interesses an der Regulierung der KI, das durch die Bletchley-Erklärung und ähnliche Initiativen signalisiert wird, bietet sich der EU hier eine einzigartige Gelegenheit.

Alle Akteure – von Wirtschaft, über Politik bis hin Zivilgesellschaft – sind nun aufgerufen, sich intensiv mit den kommenden Chancen und Risiken von Künstlicher Intelligenz, basierend auf den Leitplanken des AI Acts auseinanderzusetzen, um bei seinem vollständigen Inkrafttreten gerüstet zu sein. Neben Rechts- und Complianceberatungen dürften ethische Sensibilisierungs-

maßnahmen in Bezug auf KI zunehmend an Bedeutung gewinnen werden. Es geht darum Unternehmen bei der Integration ethischer Prinzipien in ihre KI-Anwendungen zu unterstützen. Dies könnte die Entwicklung von Ethikrichtlinien und -standards ebenso einschließen wie die Identifizierung und Bewältigung von Risiken im Zusammenhang mit KI.

Literaturverzeichnis

Zeit (2023): AI Act. Forscher fordern Umdenken der Bundesregierung bei KI-Regulierung, URL: https://www.zeit.de/digital/2023-11/ai-act-ki-gesetz-forscher-offener-brief-bundesregierung/komplettansicht#print (aufgerufen am: 12/10/2024).

(Dieser Beitrag wurde hier erstveröffentlicht: CSR News [2023]: KI-Transparenz per Regulierung, URL: https://csr-news.net/2023/12/22/ki-transparenz-per-regulierung/ [aufgerufen am: 13/10/2024].)