Positionspaper des FBKI zur Generativen KI unter besonderer Beachtung des AI Acts und der Empfehlungen des UN AI Advisory Body

Autoren: Ralf Möller, Ulrich Furbach, Katharina Morik und Simone Rehm
Weitere Mitglieder der Schreibgruppe: Christina Claß, Stefan Hildebrand, Tilmann Michaeli und Alexander Steen

In jüngster Zeit gibt es mehrere einflussreiche Initiativen, die anstreben, die Nutzung von Forschungsergebnissen der Wissenschaft der Künstlichen Intelligenz (KI) in Anwendungen normativ zu regeln. Die jeweiligen Ansätze zur Regulierung von sog. „KI-Systemen“ sind entweder gesetzgeberisch (Europäischer AI Act)[1], als Empfehlung konzipiert (UN AI Advisory Body Initiative)[2] oder als Leitlinien formuliert,[3] z.B. für den Einsatz von KI-Technologien in Schulen und Universitäten[4]. Insbesondere sind Ergebnisse der Forschung zur sogenannten generativen KI in den Fokus der aktuellen Diskussion gerückt. Mit generativen KI-Systemen sollen – so sei zunächst einmal grob umrissen – die jüngst auf der Basis großer Sprachmodelle bzw. großer multimedialer Modelle entwickelten KI-Systeme bezeichnet werden, die Sprache, Bilder, Videos oder auch allgemeine Daten und Programme, also für Menschen direkt erfassbare Ergebnisse, generieren. Bekannte generative KI-Systeme sind ChatGPT, DALL-E, Gemini, Meta AI, Perplexity.AI, um nur einige Systeme zu nennen. Bekannte (multimodale) Sprachmodelle, die als Module hinter den KI-Systemen stehen, sind GPT-4o, Llama 3, Mistral, Palm 2 (oder deren jeweilige Vorgänger) oder auch die im Projekt OpenGPT-X[5] u.a. von der deutschen Firma ALEPH ALPHA entwickelten Technologien.[6]

Im Folgenden beleuchtet die Gesellschaft für Informatik (GI) in enger Zusammenarbeit mit dem Fachbereich Künstliche Intelligenz der GI (FBKI) die aktuellen Forschungsthemen zur generativen KI und identifiziert verschiedene Ansatzpunkte für eine Regulierung der Nutzung generativer KI. Es werden auch Perspektiven für neue

[1] https://artificialintelligenceact.eu

[2] https://www.un.org/en/ai-advisory-body

[3] Die Gesellschaft für Informatik (GI) hat sich 1994 ethische Leitlinien für die Systementwicklung gegeben, die 2018 in ihrer dritten Fassung von den GI-Mitgliedern angenommen wurden. Auch wenn diese Leitlinien Künstliche Intelligenz nicht direkt nennen, beziehen sie sich generell auf die Entwicklung und Nutzung von IT-Systemen, zu denen auch KI-Systeme gerechnet werden. In der Präambel der Leitlinien steht: „Die GI-Mitglieder fühlen sich insbesondere dazu verpflichtet, die Menschenwürde zu achten und zu schützen.“ Insbesondere die „informationelle Selbstbestimmung“ wird hervorgehoben. Hier steht auch: „Die GI-Mitglieder treten dafür ein, dass Organisationsstrukturen frei von Diskriminierung sind, und berücksichtigen bei Entwurf, Herstellung, Betrieb und Verwendung von IT-Systemen die unterschiedlichen Bedürfnisse und die Diversität der Menschen.“

[4] https://www.bfb.org/handlungsempfehlungen-laender-ki
https://li.hamburg.de/ki

[5] https://opengpt-x.de/

[6] https://aleph-alpha.com/

Forschungsprogramme zur Unterstützung der Stärken und zur Abschwächung potenzieller Probleme generativer KI-Systeme entwickelt. Moral-philosophische Fragestellungen liegen außerhalb des Fokus der vorliegenden Betrachtungen.[7]

Die **zentrale Aussage** der vorliegenden Betrachtung ist, dass es weder zielführend erscheint noch global möglich sein wird, die Technik in ihrer Leistungsfähigkeit zu beschränken oder die Architektur eines Systems zur Grundlage der Regulierung zu machen. Stattdessen muss der **soziale Kontext der Nutzung eines KI-Systems durch Menschen analysiert und systematisch bewertet werden**.

Die Betrachtung geht konform mit Begriffen aus dem AI Act und den Empfehlungen des UN Advisory Body (siehe unten), betrachtet aber den sozialen Nutzungskontext und die Interaktion von Menschen (Nutzern) mit KI-Systemen wesentlich differenzierter.

Was untersucht die Wissenschaft der KI?

Langfristiges Forschungsziel der KI ist es, die Entwicklung flexibler Systeme zu ermöglichen, die explizit oder implizit gegebene Aufgaben mithilfe von intern aufgebauten Modellen über die Umgebung (die „Welt“) bestmöglich lösen. Die besondere Herausforderung ist dadurch gegeben, dass Aufgabenbeschreibungen abstrakter Natur sind und nicht direkt Algorithmen bzw. Programme darstellen.[8]

Die gerade benannten flexiblen Systeme, die übergebene Aufgabenbeschreibungen bearbeiten, werden **Agenten** im Sinne von stellvertretend handelnden, algorithmischen Akteuren genannt. Zur Bearbeitung von übermittelten Aufgabenstellungen müssen Agenten in der Lage sein,

(i) Wahrnehmungen aus der Umgebung angemessen zu interpretieren und
(ii) reaktiv Aktionen in der Welt zu bestimmen und auszuführen,
(iii) auf Rückmeldungen aus der Umwelt angemessen zu reagieren sowie
(iv) vorauszuschauen, um Änderungen in der Umgebung zu antizipieren.

Der Begriff Agent wird mit GPT-5 in Zukunft noch deutlich größere Verbreitung finden[9] [10]. Wir betrachten einige Beispiele, um die genannten Konzepte zu erläutern. Ein zukünftiger Agent, also vielleicht ein humanoider Roboter, wird an einem Tag explizit mit natürlicher Sprache instruiert, mit einer Frau, die im Pflegeheim teilnahmslos am Tisch sitzt, ein

[7] Vergleiche aber “Ethics Guidelines on Artificial Intelligence” (https://www.europarl.europa.eu/cmsdata/196377/AI%20HLEG_Ethics%20Guidelines%20for%20Trustworthy%20AI.pdf) und
“Policy and Investment Recommendations” (https://digital-strategy.ec.europa.eu/en/library/policy-and-investment-recommendations-trustworthy-artificial-intelligence).

[8] Neue Aufgabenbeschreibungen stellen also keine Reprogrammierung und auch keine Aktualisierung von Programmen dar. Auch ein KI-System mit einer festen Aufgabenbeschreibung wäre denkbar. Die Grenze von KI-Systemen zu Standardsoftware ist in diesem Falle fließend, und es greifen (auch) hier die Leitlinien der GI (https://gi.de/fileadmin/GI/Allgemein/PDF/GI_Ethische_Leitlinien_2018.pdf).

[9] https://chatopenai.de/gpt-5/

[10] https://www.youtube.com/watch?v=4Qz4GfvjGLY

anregendes Gespräch zu führen. An einem anderen bekommt er die Aufgabenbeschreibung zu melden, wenn ein alter Mann z.B. wider besseres Wissen eine Treppe verwendet und so einen Sturz riskiert. Nicht in jedem Fall muss eine Aufgabenbeschreibung durch natürliche Sprache formuliert werden. Ein Agent kann auch implizit instruiert werden, z.B. mit Bewegungen eines Exoskeletts, so dass direkt aufgeprägte menschliche Bewegungen, aber auch Kamera-basierte Umgebungsdaten kontinuierlich in dedizierte Aufgabenbeschreibungen zur menschlichen Unterstützung übersetzt und dann entsprechend umsetzt werden, um z.B. Pflegende bei bestimmten Arbeiten am Patienten zu unterstützen. Handlungen von Agenten zum Nutzen und zum Wohle aller Teilnehmer in einem sozialen Kontext können also sowohl physikalischer Art sein als auch kommunikativer Art, was die Generierung von Programmen einschließt. Das bekannte System ChatGPT der Firma OpenAI stellt einen Agenten dar, der einen Dialog mit einem Menschen (dem Nutzer) führen kann und dessen kommunikative Handlungen den mit dem System interagierenden Menschen wiederum in seinem Handeln beeinflussen. ChatGPT und der Nutzer bilden also einen sozialen Kontext, gleichermaßen wie ein humanoider Roboter im Altenheim einen sozialen Kontext mit den jeweils beteiligten und verantwortlichen Personen bildet. Um langfristig effektiv zu handeln, müssen Aufgabenbeschreibungen (als Teil der Wahrnehmungen) in einem sozialen Kontext in der Interaktion mit Menschen sinnvollerweise von den Agenten selbst immer wieder neu interpretiert werden, möglicherweise unter Ausnutzung von Rückmeldungen von Menschen im Interaktionskontext.[11]

Einen sozialen Kontext, bestehend aus Agenten und Menschen mit ihren Wechselwirkungen, wollen wir einen **sozialen Mechanismus** nennen. Ein **KI-System** fassen wir als (multiple) Agenten in einem sozialen Mechanismus auf. Die Wissenschaft der KI untersucht soziale Mechanismen.

Die Wissenschaft der KI untersucht Multi-Agenten-Systeme sowie Systeme, in denen Menschen und (multiple) Agenten interagieren, und in diesem Rahmen wurde in der KI schon sehr früh der Begriff des Mechanismus geprägt, allerdings primär bezogen auf das Zusammenwirken von Agenten.[12] [13] Agenten müssen innerhalb eines sozialen Mechanismus im Allgemeinen über die möglichen Handlungen anderer Agenten und

[11] Es wird als vollkommen ehrenwertes KI-Projekt angesehen, auch nur ein kleines Teilmodul eines sehr speziellen Agenten zu erforschen, z.B. ein Modul zur Gesichtserkennung. Es muss in der entsprechenden Publikation der Ergebnisse auch nicht ein Agent usw. erwähnt werden. Wenn man aber nun zum Beispiel ein Verfahren zur Gesichtserkennung einfach von der Stange nimmt, in eine ganz normale Maschine einbaut (Wer darf die Maschine einschalten?) und die Entwicklung der Maschine nun z.B. im Maschinenbau veröffentlicht und das Ganze dann KI-System genannt wird, dann verkommt die KI zu einem Bauchladen von Methoden, die man irgendwie verwenden kann, um dann ein ganz allgemeines Projekt zu einem KI-Projekt zu machen, sofern es opportun ist. Die gemeinhin häufig verwendete Redeweise „Die Maschine enthält eine KI, die von den Entwicklern der Maschine trainiert wurde." muss als sinnfrei angesehen werden.

[12] Rosenschein, J. S., & Genesereth, M. R. (1985). Deals among rational agents. In Proceedings of the 9th International Joint Conference on Artificial Intelligence (Vol. 1, pp. 91-99). Morgan Kaufmann Publishers Inc.

[13] Die besondere Bedeutung der Analyse von Mechanismen wird im von Thomas Malsch geprägten Wissenschaftsgebiet der Sozionik unterstrichen. Vgl. Thomas Malsch (Hg.): Sozionik. Soziologische Ansichten über künstliche Sozialität. Edition Sigma, Berlin 1998.

Menschen schlussfolgern und diese möglichen Handlungen anderer ggf. aus der eigenen Sicht bewerten.

Die Wissenschaft der KI hat bezüglich des hier beleuchteten Untersuchungsgegenstands aus technischer Sicht in jüngster Zeit unter Anderem in Hinblick auf die Entwicklung von leistungsfähigen Agenten (siehe die aktuelle Diskussion um GPT-5) wieder einmal bemerkenswerte Erfolge erzielt, was zu einem großen Presseecho geführt hat.

Warum ist KI in aller Munde?

KI ist in aller Munde, weil einerseits Agenten, wie z.B. ChatGPT, niederschwellig zur Interaktion im Web zur Verfügung stehen und ohne besondere Vorbereitung für die Menschen bedeutsame Aufgabenbeschreibungen bearbeiten können. Andererseits bekommen Entwickler leistungsfähige, sogenannte vortrainierte Modelle an die Hand, die zur Realisierung von Agenten entwickelt wurden und die zum Teil im Web frei oder gegen Gebühr zur Nutzung in Standard-Softwareprodukten bereitstehen. Den KI-Boom haben die durch US-Firmen entwickelten leistungsfähigen vortrainierten Modelle oder – darauf aufbauend – von den Firmen im Web angebotenen Agenten ausgelöst, wobei neue Techniken des maschinellen Lernens und sehr große Datenmengen aus dem Web verwendet wurden. Die aktuelle Entwicklung der umfassenden industriellen Verwendung von Agenten bzw. von vortrainierten Modellen, die auch die Entscheidungen zur Forschungsförderung in Deutschland beeinflussen könnte, ist kein Zufall.

Es konnte gezeigt werden,[14] dass mit auf Alltagsgegenständen vortrainierten Bildverarbeitungsmodellen auch eine Klassifikation von Bildern oder eine Objekterkennung und -positionierung in Bildern möglich wird, die nicht direkt mit den zum Vortrainieren verwendeten Bildern von Alltagsgegenständen verbunden sind, und zwar z.B. mit hoher Genauigkeit auch bei verhältnismäßig kleinen Mengen an Trainingsdaten. Diese wesentliche Eigenschaft wurde jüngst noch erweitert, indem neue, von den Menschen für ihre Unterhaltung oder Arbeit als sinnvoll erachtete mediale Inhalte erzeugt werden können. Diese Entwicklung wird als generatives Verhalten (von Agenten) oder kurz **generative KI** im Sinne einer neuen Technologie bezeichnet.

Beispiele für generatives Verhalten zur Umsetzung von Aufgabenbeschreibungen sind die Übersetzung oder Zusammenfassung von Texten, die Schritt-für-Schritt-Beschreibung von Problemlösungen für natürlichsprachlich gegebene Rätsel oder auch die Generierung von Bildern oder Videos. Es ist allerdings zu konstatieren, dass bei den genannten Aufgabenbeschreibungen, die den Agenten übergeben werden, im Allgemeinen kein Korrektheitsbegriff gegeben ist, sondern die „Leistung“ von Agenten über den Begriff des von außen als „intelligent“ erscheinenden Verhaltens der Agenten erfasst wird. Der Intelligenzbegriff in der KI bezieht sich auf das Handeln von Agenten.

[14] Siehe zum Beispiel: A. Razavian, H. Azizpour, J. Sullivan and S. Carlsson, "CNN Features Off-the-Shelf: An Astounding Baseline for Recognition," in 2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Columbus, OH, USA, 2014 pp. 512-519. https://doi.ieeecomputersociety.org/10.1109/CVPRW.2014.131

Der **Intelligenzbegriff** des in Grenzen[15] lokal optimalen Handelns von Agenten in Bezug auf unsichere Interpretationen von Aufgabenbeschreibungen und den Präferenzen anderer in der Umgebung (dem sozialen Mechanismus) kennzeichnet die Wissenschaft der KI und bestimmt dadurch wichtige Herausforderungen in der KI-Forschung.[16]

Der Volksmund spricht häufig von „einer KI“ im Sinne eines intelligent Handelnden und meint einen Agenten.

Erfolge und Herausforderungen in der KI-Forschung

Aufgrund der Komplexität oder des Fehlens einer Spezifikation von Korrektheitsbedingungen des Handels von Agenten oder von Ergebnissen von Entscheidungs- bzw. Berechnungsproblemen über vortrainierte Modelle, sind neue, wichtige Strategien zum Testen von KI-Systemen entwickelt worden.[17]

Es wird deutlich, dass das Thema **Testen** von Agenten, bei aller Hochachtung für die schon erzielten Erfolge, in bestimmten Mechanismen noch deutlich intensiver als bisher erforscht werden muss (mit Themen wie z.B. Vermeidung von Bias-Effekten,[18] Fairness, Unterbindung von Hassreden, Einhaltung von Persönlichkeitsrechten, Schutz geistigen Eigentums, ...).

Das Testen der „Angemessenheit“ von KI-Systemen in einem Anwendungskontext (also einem sozialen Mechanismus, wenn man es technisch ausdrücken möchte) ist sicher weit komplexer als z.B. das Testen der Generalisierungsfähigkeit von einzelnen (statistischen) Schätzern oder Generatoren, die durch Techniken des maschinellen Lernens gewonnen wurden. Es wurden mit dem „Bonner KI-Prüfkatalog“ auch Tests z.B. auf Vertrauenswürdigkeit vorgeschlagen.[19] Aufgrund der prinzipiellen Beschränkung von Tests (gefundene Fehler können behoben werden, aber grundsätzlich kann Fehlerfreiheit nicht garantiert werden), ist auch die forensisch motivierte Erklärbarkeit mit angemessenen Strategien zur Reportgenerierung ein wichtiges Forschungsthema.[20] Erklärbarkeit des Verhaltens von Agenten zur „Laufzeit“ kann in Einzelfällen auch für Menschen in einem sozialen Mechanismus bedeutsam sein (neben dem sicherlich zu Recht geforderten Mensch-kompatiblen und -bezogenen Verhalten von Agenten an

[15] Agenten müssen in vielen Anwendungen ihre Entscheidungen für Handlungen in engem Zeitrahmen treffen, so dass z.B. Anytime-Algorithmen eingesetzt werden müssen, weil die Komplexitätstheorie der Informatik auch für Agenten greift.

[16] https://www.soziopolis.de/zwischen-macht-und-mythos.html

[17] Es gibt bereits erste Ansätze für Tests der Robustheit die Bibliotheken CleverHans (https://github.com/cleverhans-lab/cleverhans) und RobustBench (https://robustbench.github.io), für den Energieverbrauch die sog. Care Labels, für die Fairness die Bibliotheken Fairness Indicator in TensorFlow (https://www.tensorflow.org/tfx/guide/fairness_indicators) und Fairness Tree in Python (https://dssg.github.io/fairness_tutorial/).

[18] Dino Pedreschi, S.Ruggieri, F. Turini (2008) Discrimination-aware data mining. In: KDD 2008.

[19] https://www.iais.fraunhofer.de/de/forschung/kuenstliche-intelligenz/ki-pruefkatalog.html

[20] Raphael Fischer, Thomas Liebig, Katharina Morik (2024) Towards More Sustainable and Trustworthy Reporting in Machine Learning. Data Mining and KDD Journal.

sich)[21] und muss deutlicher auf den gesamten Mechanismus Bezug nehmen, so dass sich auch hier neue Forschungsziele ergeben. Mit den jüngst gewonnenen neuen Erkenntnissen im Bereich des Testens (siehe oben) sowie den Arbeit zur Gewährleistung der Interpretierbarkeit gelernter Modelle[22] wird die Analyse von Mechanismen sicherlich einfacher, kann aber kaum gänzlich eliminiert werden.

Ob durch die Kommunikation für die Nutzer von neuen Softwaresystemen, die auf Basis von generativer KI-Modelle entwickelt wurden oder für die Nutzer von Agenten eine Gefahr besteht, hängt vom sozialen Mechanismus ab und wird derzeit in der Öffentlichkeit intensiv diskutiert. Bestrebungen zur Regulierung der Entwicklung und des Einsatzes von intelligenten Systemen versuchen, Gefahren für Menschen abzuwehren, die sich aus dem Umgang mit (generativen) KI-Systemen ergeben könnten.

Regulierungsbestrebungen

In dem im Dezember 2023 veröffentlichten Zwischenbericht der Vereinten Nationen „Governing AI for Humanity“, verfasst vom UN AI Advisory Body, wird auf die OECD-Definition eines KI-Systems Bezug genommen.[23] In dieser Definition wird die oben aufgeführte Akteurs- bzw. Agenten-basierte Sicht auf ein KI-System zugrunde gelegt, und es wird nahegelegt, die Gestaltung der Implementierung der Agentenarchitektur (also den internen Aufbau eines Agenten) in geeigneter Weise zu reglementieren.[24] Für eine Regulierung in diesem Kontext ist es bedeutsam, dass ein Agent ein kontextbezogener Akteur ist, der seine Umgebung wahrnimmt, daraus Informationen über die Umgebung entnimmt und auch Aufgabenbeschreibungen für sich selbst ableitet oder anpasst. Auf die Verwendung eines Korrektheitsbegriffs des Handelns eines Agenten wird auch in der OECD-Definition und damit in der Empfehlung des UN Advisory Bodys verzichtet.[25] Es wird vorgeschlagen, dass die Struktur einer Agentenarchitektur in geeigneter Weise geregelt werden sollte. Eine Anwendung von Agenten-basierten KI-Systemarchitekturen allein und ohne Kontextbetrachtungen zu reglementieren, erscheint aus Sicht der KI-Forschung nicht sinnvoll, wie im Folgenden weiter ausgeführt wird.

---

[21] Kambhampati, S. (2022). Human-Aware AI. MIT Press.

[22] Wojciech Samek, Gregoire Montavon, Andrea Vedaldi, Lars Kai Hansen, Klaus-Robert Müller (Hg.) (2019) Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. Springer.

[23] https://oecd.ai/en/wonk/ai-system-definition-update

[24] Trotz der Tatsache, dass Stuart Russell ein Co-Autor der OECD-Definition des Begriffs KI-System ist, wird in der 2023 aktualisierten Definition auf Einsichten aus seinem Buch “Human Compatible” (Russell, S. J. (2019). Human Compatible: Artificial Intelligence and the Problem of Control. Viking) mit Themen wie Unsicherheit des Agenten bzgl. der angenommenen Ziele, beweisbar-nützliches Verhalten oder Assistenz-basierte Spieltheorie in der Definition für ein KI-System nicht eingegangen. Ob für Agenten, die die Einsichten aus “Human Compatible” beinhalten, die vorgeschlagenen Regularien nicht gelten sollen, bleibt unklar.

[25] Im gesellschaftlichen Umfeld spüren die Menschen genau, dass die Systeme, mit denen sie interagieren, nicht einfache Programme oder direkt-manipulative Oberflächen sind, und sie verwenden daher den Begriff “eine KI”, meinen aber genau einen Agenten damit, ohne den im wissenschaftlichen Kontext sinnvollerweise nüchternen Fachterminus “Agent” zu verwenden.

Der europäische AI Act ist der weltweit erste umfassende Rechtsrahmen für KI-Systeme, der darauf abzielt, den vertrauenswürdigen Einsatz von KI-Technologie in Europa zu fördern, und klare Anforderungen für KI-Entwickler und Verantwortliche für den Einsatz von KI-Systemen festlegt. Bürger haben das Recht, Beschwerden über KI-Systeme einzureichen und Erklärungen zu erhalten, insbesondere bei KI-Systemen mit sog. hohem Risiko für die beteiligten Menschen. Hochrisikosysteme nach der Klassifikation des europäischen AI Acts sind KI-Systeme, die ein erhebliches Risiko für die Gesundheit, Sicherheit oder Grundrechte darstellen. Diese Systeme werden in verschiedene Kategorien eingeteilt und unterliegen strengen Compliance-Anforderungen gemäß den Bestimmungen des AI Acts. Anbieter solcher Hochrisikosysteme müssen Konformitätsbewertungen durchführen und sich an spezifische Vorgaben halten, um die Einhaltung der Regulierungen sicherzustellen.[26] Für Hochrisikosysteme bestehen größere Anforderungen an Transparenz und Zugang. Der Anwendungsbereich des AI Act erstreckt sich auf Anbieter, die KI-Systeme in der EU in Verkehr bringen oder betreiben, sowie auf Nutzer von KI-Systemen innerhalb der Union. Sowohl juristische Personen als auch natürliche Personen sind betroffen.[27]

Wer allerdings ein System als risikoreich für den Menschen einstuft, muss die sozialen Mechanismen berücksichtigen, in denen Menschen mit Systemen und auch untereinander interagieren. Was in dem einen Mechanismus ein hohes Risiko darstellt, kann in einem anderen Mechanismus tolerierbar sein. Zum Beispiel können die Probleme von Verfahren zur Gesichtserkennung in der Breite sehr bedeutsam sein (Diskriminierungseffekte), in einem speziellen industriellen Anwendungskontext (Darf ein Arbeiter eine Maschine einschalten?) müssen Probleme dieser Art aber nicht unbedingt zutage treten.

Die Definition eines KI-Systems gemäß dem AI Act fußt auf einem Begriff der „Inferenz“, um KI-Systeme von klassischer Software abzugrenzen.[28] KI-Systeme bringen Ergebnisse wie Inhalte, Vorhersagen, Empfehlungen oder Entscheidungen hervor, die das Umfeld beeinflussen, mit dem sie interagieren, und legen damit den Rahmen fest, innerhalb dessen KI-Systeme im Kontext des AI Acts betrachtet und reguliert werden. Also liegt auch dieser Definition des Begriffs „KI-System“ im Prinzip der Gedanke eines Agenten zugrunde, der in einer Umgebung, einem sozialen Kontext, wirkt. Aus der Betrachtung der Regulierungsbestrebungen ergibt sich also, dass sich die aktuelle öffentliche Diskussion über (generative) KI um oben eingeführte Begriffe wie „sozialer Mechanismus“[29] und

[26] Die rechtlichen Instrumente, mit denen entschieden wird, ob bei einem bestimmten Mechanismus ein hohes Risiko für den Menschen besteht, sind nach wie vor unzureichend definiert.

[27] Bei Verstößen gegen die Verordnungen des AI Act können Bußgelder von bis zu 35 Millionen Euro oder sieben Prozent des weltweiten Jahresumsatzes des Unternehmens verhängt werden, um die Einhaltung der Regulierungen sicherzustellen.

[28] Der Begriff der “Inferenz” ist im Rahmen des AI Acts folgendermaßen definiert: “This capability to infer refers to the process of obtaining the outputs, such as predictions, content, recommendations, or decisions, which can influence physical and virtual environments, and to a capability of AI systems to derive models or algorithms, or both, from inputs or data.” (a.a.O. Absatz (12)).

[29] Social Mechanisms: An Analytical Approach to Social Theory. Studies in Rationality and Social Change, Cambridge University Press (1998)

„Agent“ drehen muss (ob ein Agent derzeit nun gemeinhin als „eine KI“ bezeichnet wird oder nicht).

Die Entwicklung sozialer Mechanismen erfordert die zielgerichtete Untersuchung des Zusammenwirkens von Agenten und Menschen. Mechanismen müssen mit Bedacht entwickelt werden, um das nur lokal optimale, aber möglicherweise global eigennützige Verhalten der Agenten (und auch der Menschen), die in einem Mechanismus interagieren, entsprechend mit Anreizen so zu lenken, so dass eine vorgesehene globale „soziale Funktion“ zielgerichtet berechnet wird.[30]

Mit zunehmender Leistungsfähigkeit von Agenten agieren diese auch direkt mit Menschen, und es entsteht die **Notwendigkeit, soziale Mechanismen zu untersuchen**. Mit den neuen Entwicklungen der generativen KI kommt hinzu, dass Menschen in nicht wenigen sozialen Mechanismen Gefahr laufen, Agenten nicht von Menschen unterscheiden zu können (vgl. die Debatte zu DeepFakes)[31], und darin liegt derzeit die Brisanz des gesamten Technologiefeldes „Generative KI“ – auch gerade wegen der Tatsache, dass interne Ziele und Verhaltensweisen von KI-Systemen mit menschlichen Werten und Absichten nicht (immer) übereinstimmen.[32]

Nun mag es für einige durchaus befremdlich sein, dass man einmal aus der Perspektive des hier vertretenen nicht Mensch-zentrierten Intelligenzbegriffs und über soziale Mechanismen und dem Zusammenwirken von Menschen und Agenten greifbar zu machen versucht, worum es in der Wissenschaft der KI überhaupt geht. Mittlerweile gibt es jedoch auch vielfache Ansätze aus der Ökonomie, die KI nicht mit dem Maßstab der (Human Level-)Intelligenz, sondern mit ihrem Grad der Nützlichkeit für die Wirtschaft und die Gesellschaft an sich messen wollen[33]. Es wird dort argumentiert, dass nicht mehr Tests, wie der Turing-Test, die ein KI-System isoliert betrachten, verwendet werden sollen, sondern das Zusammenwirken in einer komplexen Struktur untersucht werden muss.[34] [35] Auch die vielfältigen Diskussionen in der Öffentlichkeit über die Auswirkungen

[30] Mechanismus-Design wird in der vierten Ausgabe zwar auch in dem für die KI sehr bedeutsamen Buch „Artificial Intelligence: A Modern Approach“ von S. Russell und P. Norvig (Pearson Verlag) richtigerweise systematisch erwähnt, die Darstellung ist aber noch zu stark auf das Zusammenwirken von nur artifiziellen Agenten bezogen. Rao Kambhampati u.a. arbeiteten hier weiter, gehen aber auch noch nicht weit genug, siehe Sarath Sreedharan, Anagha Kulkarni, Subbarao Kambhampati (2022) Explainable Human-AI Interaction: A Planning Perspective Sreedharan, Morgan and Claypool Publishers.

[31] https://www.bsi.bund.de/DE/Themen/Unternehmen-und-Organisationen/Informationen-und-Empfehlungen/Kuenstliche-Intelligenz/Deepfakes/deepfakes_node.html

[32] Koster, R., Balaguer, J., Tacchetti, A., Weinstein, A., Zhu, T., Hauser, O., Williams, D., & Lucy, L. (2022). Human-centered mechanism design with Democratic AI. Nature Human Behaviour, 6(10), 1383-1393.

[33] Acemoglu, D., Johnson, S.: Power and Progress. Basic Books (2023)

[34] Star, S.L.: Die Struktur schlecht strukturierter Lösungen. Grenzobjekte und heterogenes verteiltes Problemlösen. In: Star, S.L. (ed.) Grenzobjekte und Medienforschung, pp. 131–150. transcript Verlag, Bielefeld (2017).

[35] Der auf den Turing-Test zurückgehende veraltete Intelligenzbegriff, der auf Simulation des Menschen basiert, verleitet leider immer noch viele dazu, ein simples Programm als „intelligent“ zu bezeichnen, wenn es etwas leistet, das einen sehr kleinen Teil davon ausmacht, was Menschen zu leisten im Stande sind (und Programme eben vorher nicht). Dieser Intelligenzbegriff hat sich nicht bewährt und führt zu Paradoxien, denn dann wird allzu leichtfertig eine einfache (mathematisch wohldefinierte) Funktion, die über Techniken des maschinellen Lernens aus Daten generiert wurde, als intelligent oder eben als „eine

von KI auf den Arbeitsmarkt und die Ängste vor einer möglichen Vorherrschaft von KI-Systemen zeigen, dass es dringend einer interdisziplinären und systemorientierten Behandlung des Themas bedarf.

Der Begriff des sozialen Mechanismus als Ansatzpunkt zur Regulierung

Die Gesellschaft für Informatik vertritt die These, dass Regulierung am Begriff des sozialen Mechanismus ansetzen muss, dass also AI Act und UN-Ansätze weitergedacht werden müssen. Sinnvoll regulieren kann man, dass ein für die Anwendung vorgesehener Mechanismus, in dem Agenten und Menschen interagieren, „hinreichend gut" auf unerwünschte Effekte und Risiken für die Menschen hin untersucht wird. Unerwünschte Effekte können sein, dass im Mittel Menschen zwar gut „wegkommen", z.B. in einer Situation relativ viele Zuschüsse bekommen, aber es nicht ausgeschlossen werden kann, dass Einzelne leer ausgehen und dadurch möglicherweise in wirtschaftliche Not geraten. Es müssen bei einer hinreichend guten Analyse oder schon bei der Konstruktion von Mechanismen Argumente dafür gefunden werden, dass die gewünschten Effekte eintreten und ungewollte eben nicht (dass also z.B. Menschen nicht (ungewollt) zu Schaden kommen), was auf technischer Ebene so formuliert wird, dass eine bestimmte soziale Funktion auch wirklich in der Situation realisiert wird.[36] [37] [38]

Man könnte durchaus argumentieren, dass vorhandene Technologie an sich schon gefährlich sein kann, weil Menschen verführt werden, sie zum Schaden anderer zu nutzen.[39] Im Vergleich zur Nukleartechnologie liegt aber bezüglich Missbrauchsmöglichkeiten eine nicht vergleichbare Situation vor. Wenn allerdings in einem konkreten Fall keine guten Argumente für das „Funktionieren" eines in der Praxis möglicherweise zu realisierenden Mechanismus vorgebracht werden können, wenn also die mit dem Mechanismus assoziierte soziale Funktion für die beteiligten Menschen auch nur potenziell (hohe) Risiken oder (starke) Nachteile für Einzelne birgt, dann kann die Verwendung des Mechanismus entsprechend eingeschränkt oder sogar verboten werden. Es kann aber sein, dass bestimmte KI-Systeme, die nach AI Act bzw. nach den Empfehlungen des UN AI Advisory Bodys mit einem Bann zu versehen sind, sehr wohl in bestimmten sozialen Mechanismen ohne Risiko eingesetzt werden können, während andere durch als niedrig eingestuftes Risiko sogar durch eben dieses Raster fallen, obwohl es soziale Mechanismen gibt, in denen dennoch für Menschen nachteilige Effekte auftreten können.

KI" bezeichnet (z.B. eine Funktion für die Klassifikation eines Bildes als Darstellung einer krankhaften oder einer gesunden Struktur).

[36] Vohra, R. V. (2011). Mechanism Design: A Linear Programming Approach. Cambridge University Press.

[37] Börgers, T. (2015). An introduction to the theory of mechanism design. Oxford University Press.

[38] Fickinger, A., Zhuang, S., Critch, A., Hadfield-Menell, D., & Russell, S. (2020). Multi-Principal Assistance Games: Definition and Collegial Mechanisms. arXiv preprint arXiv:2012.14536.

[39] . In diesem Zusammenhang ist auch die Diskussion um das Off-Switch-Problem wichtig.

Der Mechanismusbegriff der KI ist der **Schlüssel zur Regulierung der Nutzung (und damit auch der Entwicklung) von KI-Systemen**. Die Analyse eines Mechanismus, der zur Anwendung kommen soll, lässt sich vorschreiben und vielleicht sogar von unabhängigen (internationalen) Prüforganisationen durchführen. KI-Systeme als Technologie an sich zu regulieren ist aus Sicht der Gesellschaft für Informatik nicht zielführend.

Desiderata

Begrifflich konform mit, aber anders ausgerichtet als der AI Act wird eine Art Mechanismus-Zertifizierungsstelle („Mechanismus-TÜV"), also evidenzbasierte, unabhängige Voten für die Gestaltung von Mechanismen in speziellen Anwendungsfällen, für sinnvoll erachtet. Der zusätzliche Aufwand (geleistet ggf. durch neue Berufsfelder) wird durch den entstehenden Mehrwert für die Gesellschaft kompensiert, wenn die mit der Idee der Analyse von sozialen Mechanismen einhergehenden sozialen und technischen Forschungsfragen erfolgreich bearbeitet sind, und zwar in direktem Anschluss an die erfolgreichen vorigen Forschungsarbeiten in Deutschland, sei es zu Formalismen, Trainings- und Testtechnologie oder auch zu speziellen Themen, wie z.B. Erklärungsgenerierung für das Verhalten von Agenten (z.B. zur Forensik im nicht auszuschließenden Fehlerfall). Mechanismusanalyse muss mittels interdisziplinärer Zusammenarbeit erforscht werden.[40] [41] Da sich die Randbedingungen der Anwendung von Systemen in der Realität ändern können, ist in einigen Anwendungskontexten sogar eine periodisch wiederholte Mechanismus-Analyse einzufordern.

Eine Mechanismus-Analyse kann die Einsicht erbringen, dass zunächst einige Voraussetzungen geschaffen werden müssen, bevor ein Mechanismus etabliert werden kann. So kann zum Beispiel eine vorige Ausbildung der menschlichen Mechanismus-Teilnehmer als notwendig erachtet werden. Bei Anwendungen von generativer KI in offenen Mechanismen, also Mechanismen, in denen die Identität der beteiligten Menschen nicht a priori bekannt ist (wie z.B. bei ChatGPT und seinen Nutzern), ist das Fehlen eines Korrektheitsbegriffs besonders bedeutsam, und weitere interdisziplinäre Forschungsarbeiten sind notwendig, um diese Form von Mechanismen besser zu verstehen und zu analysieren. Forschungsarbeiten zur direkten Integration von moralisch-normativen Aspekten in die intrinsische Handlungsplanung von Agenten in die jeweils lokalen internen Nützlichkeitsbewertungsmaße von Agenten können das Design von sozialen Mechanismen eventuell vereinfachen.[42] Aber auch in diesem Gebiet bedarf

[40] Star, S.L.: Die Struktur schlecht strukturierter Lösungen. a.a.O.

[41] Suryanarayana, S.: Human consideration in analysis and algorithms for mechanism design. In: Baumeister, D., Rothe, J. (eds.) Multi-Agent Systems - 19th European Conference, EUMAS 2022, Proceedings. pp. 444–447. Springer (2022)

[42] Schramowski, P., Turan, C., Jentzsch, S., Rothkopf, C., and Kersting, K. (2020). The Moral Choice Machine. Frontiers in Artificial Intelligence, 3.

es weiterer intensiver Arbeit, damit die Ergebnisse der Forschung u.a. im beschriebenen Sinne zum Vorteil von Wirtschaft, Wissenschaft und Gesellschaft zur Wirkung kommen.

Insgesamt wird deutlich, dass sich Forschungsbedarf in zwei Richtungen ergibt, um Wirtschaft und Gesellschaft voranzubringen:

- Interdisziplinäre Forschung zum Design und zur Analyse von sozialen Mechanismen, in denen Menschen und intelligente Agenten auf Basis generativer Modelle interagieren, sowie Forschung zur Schulung von Menschen bei der Interaktion mit intelligenten Agenten. KI-Wissenschaftler arbeiten in diesem Fall mit Soziologen, Erziehungswissenschaftlern, Geisteswissenschaftlern usw. zusammen.

- Forschung zu im Kontext von Agenten und Mechanismen geeigneten Formalismen und entsprechenden Architekturen und Trainingstechniken zum Aufbau und Nutzen von vortrainierten Modellen, so dass Anwender nicht Formalismen studieren müssen, um Modelle erstellen und nutzen zu können, und auch auf Basis von vorhandenen kleinen Datenmengen (unter Ausnutzung der vortrainierten Modelle) Anpassungen vornehmen können (Few-Shot Learning, Fine-Tuning, Prompt-Generierung, Destillation und Kalibrierung von anwendungsspezifischen Modellen aus großen vortrainierten Modellen). Auch Fragen des De-Biasings und der Privatheit sowie der Anonymisierung von Trainingsdaten spielen hier eine große Rolle.

Obwohl im Bereich des ressourcenschonenden Erstellens von (vortrainierten) Modellen und Agenten durch maschinelles Lernen schon große Erfolge erzielt wurden,[43] sind weitere Anstrengungen notwendig.[44] Weiterhin kann interaktives maschinelles Lernen[45] [46] der Agenten von den Menschen in einem sozialen Mechanismus helfen, bessere Modelle (und damit Agenten) zu gewinnen, was gerade auch in Deutschland erforscht wird.[47] Auch das algorithmisch unterstützte Erstellen von Mechanismen wird immer bedeutsamer.[48]

Es sei hier beispielhaft insbesondere noch einmal auf OpenGPT-X verweisen.[49] Auch die Öffnung der A/B-Testing-Infrastruktur der großen Online-Plattformen (z.B. von X,

[43] Der SFB 876 war der ressourcenschonenenden automatisierten Erstellung von Agenten und Mechanismen gewidmet. Die Ergebnisse sind in den drei Bänden zu Machine Learning under Resource Constraints festgehalten (https://www.degruyter.com/serial/mlrc-b/html).

[44] Vgl. z.B. https://www.spektrum.de/news/kuenstliche-intelligenz-verbraucht-fuer-den-lernprozess-unvorstellbar-viel-energie/1660246

[45] Amershi, S., Cakmak, M., Knox, W. B., & Kulesza, T. (2014). Power to the people: The role of humans in interactive machine learning. AI Magazine, 35(4), 105-120.

[46] Argall, B. D., Chernova, S., Veloso, M., & Browning, B. (2009). A survey of robot learning from demonstration. Robotics and autonomous systems, 57(5), 469-483.

[47] https://www.dfki.de/web/forschung/forschungsbereiche/interaktives-maschinelles-lernen

[48] Vgl. Nagai, Y., & Conitzer, V. (2016). Algorithmic Mechanism Design. Cambridge University Press.

[49] https://opengpt-x.de/

Amazon, Meta) für wissenschaftliche Studien ermöglicht es Forschenden, kontrollierte Experimente mit einer großen Nutzerbasis durchzuführen und so wertvolle Erkenntnisse über menschliches Verhalten, Entscheidungsfindung und die Wirksamkeit verschiedener Designs oder Funktionen zu gewinnen. Die Nutzung dieser Plattformen für die Forschung erfordert aber eine sorgfältige Planung, Kontrolle und Berücksichtigung der Privatheit bzw. Anonymisierung. Zusammenfassend lässt sich Folgendes sagen.

Forschung wird erst möglich durch zielstrebiges Propagieren von Ansätzen, die sich den Konzepten **Open Source, Open Data, Open Services** verschrieben haben.

Zusammenfassung

In diesem Aufsatz zu Themen um generative KI und die Ansatzpunkte der Regulierung des Einsatzes von KI-Systemen geht es um das *Was* aus der KI, also Agenten und Menschen in sozialen Mechanismen. Es geht nicht um das *Wie* der Realisierung eines Agenten (z.B. durch Techniken des maschinellen Lernens). Wir haben beleuchtet, dass durch den AI Act verschiedene Formen der Regulierung an einen Risiko-Begriff, der sich auf einen sozialen Mechanismus beziehen muss, gekoppelt sind. Das ist ein zentraler Schritt, und der Begriff des Agenten, von dessen Handlungen das Risiko ausgeht, ist ein wichtiges Konzept. Der Begriff des „Risikos“ bedarf allerdings weiterer Betrachtung. Derzeit sind viele Menschen auch einfach besorgt um ihre Arbeitsplätze, die z.B. durch (generative) KI-Systeme nach ihrem Empfinden verloren gehen können (Risiko des Arbeitsplatzverlustes). Allerdings kann generative KI auch zum Risiko für demokratische Strukturen werden, und weitere Risiken lassen sich aufzählen. Beim derzeitigen Stand umfasst der AI Act die für Menschen bedeutsamen, langfristigen Konsequenzen (Risiken) dieser Art nicht. Da derzeit sehr viele unter Umständen Arbeitsplatz-vernichtende normale Softwaresysteme zur Industrie-Automation, die einen aus Daten gewonnenen Schätzer oder Klassifikator enthalten, als KI-Systeme hochstilisiert werden, kommt die KI als Wissenschaft in eine ungerechtfertigte Regulierungs-Bredouille. Nur besondere KI-Techniken oder bestimmte Architekturen in KI-Systemen regulieren zu wollen und „normale“ Software eben nicht, wird daher als nicht zielführend angesehen. Wenn Regulierung auf diese Art organisiert wird, werden die Entwickler die eingesetzten Techniken einfach nicht mehr als KI-Techniken bezeichnen (und die Systeme nicht mehr als KI-Systeme). So wie die Leitlinien der GI[50] auch für KI-Systeme gelten, sollten die Überlegungen des AI Acts auch für „normale Softwaresysteme“ mit einem handelnden Einfluss in einem sozialen Kontext greifen.

Insgesamt heben wir als Anwender:innen und als Forscher:innen innerhalb der Gesellschaft für Informatik hervor, dass wir Regulierungen für die Anwendung von KI-Systemen für wichtig erachten, dass aber der Fokus der Regulierung nicht auf die KI-Systeme selbst bzw. deren interner Technologie gerichtet werden darf, sondern auf die Interaktion der KI-Systeme mit dem Menschen ausgerichtet werden muss. Aus einem solchen erweiterten Ansatz erwachsen große Chancen sowohl für die Anwendung bestehender Technologien als auch für die KI-Forschung in Deutschland. Die

[50] https://gi.de/ueber-uns/organisation/unsere-ethischen-leitlinien

hervorragenden Vorarbeiten auf dem Gebiet der KI-Forschung in Deutschland können durch eine grundlagenorientierte Forschungsförderung im Sinne der genannten Perspektiven in Zukunft noch deutlich Fahrt aufnehmen – zum Wohle von Wirtschaft, Wissenschaft und Gesellschaft.

Danksagung

Wir bedanken uns für Rückmeldungen bei den KI-Fachgruppensprechern Andreas Hotho, Özgür Özcep, Benjamin Paaßen und Diedrich Wolter sowie bei den Kolleg:innen Michael Friedrich, Antonio Krüger, Philipp Rostalski und Gesina Schwalbe.