HELMHOLTZ

Spitzenforschung für
große Herausforderungen.

Handlungsempfehlungen zum Einsatz von Künstlicher Intelligenz

Version 1.0

(von der Helmholtz-Mitgliederversammlung beschlossene Fassung vom 18.09.2024)

Inhalt

1. Präambel

Künstliche Intelligenz (KI), ihre vielfältigen Einsatzmöglichkeiten und potentiell damit verbundene Risiken beschäftigen derzeit große Teile der Gesellschaft. Anlass dafür sind insbesondere die Entwicklung und großflächige Nutzbarmachung von generativen KI-Systemen für die Text-, Video-, Bild- oder Musikerstellung (z.B. ChatGPT, Dall-E).

Bereits vor der Veröffentlichung bekannter generativen KI-Systeme war KI im beruflichen und privaten Alltag vieler integriert. Während Navigationssoftware (z.B.Google Maps) KI bei der Routenplanung nutzt, verwenden Streamingdienste (z.B. YouTube, Netflix) und Verkaufsplattformen (z.B. Amazon) KI für personalisierte Empfehlungen basierend auf Nutzerverhalten und Vorlieben. Im beruflichen Kontext kommt KI mittlerweile großflächig zum Einsatz, z.B. bei der Rechtschreib- und Grammatikprüfung (z.B. Microsoft Word, Grammarly), bei der Analyse und Datenvisualisierung (z.B. Microsoft Excel) oder bei der automatischen Hintergrundgeräuschunterdrückung (z.B. Zoom). Doch erst die umfassende Nutzbarmachung von generativen KI-Systemen, also großen KI-Modellen, die auf umfangreichen Datenmengen trainiert werden und darauf basierend neue Inhalte generieren, die kaum noch von menschlich Produziertem zu unterscheiden sind, verdeutlichte gesamtgesellschaftlich das enorme Potential von KI[1].

KI ist in der Helmholtz-Gemeinschaft ein relevanter Forschungsgegenstand und die Entwicklungen in dem Bereich daher von großer Bedeutung. So verfügt Helmholtz zum einen über selbst gehostete, große Rechenkapazitäten (z.B. JUWELS Booster, HoReKa und bald JUPITER) und stellt diese Forschenden und (externen) Nutzern zur Verfügung, um (generative) KI-Systeme zu entwickeln und anzuwenden. Zum anderen treibt Helmholtz aktiv die Entwicklung von eigenen KI-Modellen für eine Vielzahl von Forschungsbereichen voran (z.B. Helmholtz-Foundation-Modell-Initiative (HFMI)). Gleichzeitig entwickeln Forschende in Helmholtz neue KI-Methoden oder wenden KI-Technologien und -Systeme an, um mit deren Hilfe Lösungen für die großen Herausforderungen unserer Zeit zu entwickeln.

Als Europas größte Forschungsorganisation ist Helmholtz nicht nur daran gelegen, die Welt der KI durch die Entwicklung von eigenen KI-Systemen mitzugestalten, sondern auch daran, die Mitarbeitenden in allen Bereichen der Gemeinschaft zur verantwortungsvollen Nutzung zu ermutigen und zu befähigen.

Ziel dieser Leitlinien soll es daher sein, neben den Chancen durch KI-Systeme auch potentielle Risiken und Handlungsempfehlungen zu deren Minimierung aufzuzeigen. Dies soll alle Mitarbeitenden der Helmholtz-Gemeinschaft dabei unterstützen, KI-Systeme informiert, reflektiert und verantwortungsbewusst zu verwenden bzw. zu entwickeln.

2. Begriffserklärung

Es gibt bislang keine allgemein anerkannte Definition von KI und der rasante Fortschritt macht es unwahrscheinlich, dass bald ein Konsens erzielt wird. Es lassen sich aber einige Merkmale benennen, die verbreitet zur Charakterisierung von KI verwendet werden.

Künstliche Intelligenz (KI) bezeichnet den Bereich der Informatik, der sich mit der Entwicklung von Systemen und Algorithmen befasst, die es zum Ziel haben, Aufgaben für die digitale und physische Welt zu lösen.

Traditionelle KI-Systeme, welche auch als regelbasierte oder symbolische KI-Systeme bekannt sind, beruhen auf vorgegebenen Regeln und Algorithmen, um Entscheidungen zu treffen und

[1] Siehe z.B. „Science in the age of AI“ der Royal Society.

Probleme zu lösen. Diese Art der KI nutzt logische Schlussfolgerungen und explizit codierte Wissensdatenbanken, um Aufgaben zu erfüllen.

Generative KI-Systeme sind eine spezielle Form der Künstlichen Intelligenz, die über Automatisierung und Klassifizierung traditioneller KI hinausgeht. Sie produzieren neue Inhalte (Texte, Bilder, Videos, Musik usw.), indem sie Muster und Strukturen in großen Datenmengen in einem Trainingsprozess lernen und replizieren.

3. Chancen durch KI

KI-Systeme haben das Potenzial, viele Aufgaben in Forschung, Management und Verwaltung zu erleichtern.

In der Forschung ist Künstliche Intelligenz schon seit langer Zeit ein fester Bestandteil des Methodenrepertoires. Die in jüngster Zeit entwickelten generativen KI-Systeme versprechen dramatische Vereinfachungen datenwissenschaftlicher Prozesse. Insbesondere in der explorativen Phase der Forschung, in der Hypothesen generiert und getestet werden, können KI-Systeme zu einer Beschleunigung des Erkenntnisprozesses beitragen, etwa bei der Programmierung neuer Werkzeuge. Hinzu kommt, dass generative KI-Systeme die Anwendung komplizierter Analysealgorithmen vereinfachen können, etwa durch natürlichsprachliche Schnittstellen, und diese somit einer breiteren Forscherbasis zugänglich machen. Aber auch die Automatisierung wiederkehrender Aufgaben kann Effizienzen schaffen, wie zum Beispiel in der Datenaufbereitung oder bei der Generierung von Berichten. Dadurch können Forschende mehr Zeit in kreative und strategische Tätigkeiten investieren, was die Innovationskraft und wissenschaftliche Produktivität erheblich steigert.

(Generative) KI-Systeme bieten aber auch viele Anwendungsmöglichkeiten für die Verwaltung und das Management. Von Chatbots, die interaktiv den Zugang zu Wissen in Datenbanken oder Cloudsystemen ermöglichen, bis hin zum Verfassen von Besprechungsnotizen, Stellenausschreibungen oder Texte für vielfältige andere Kontexte – große Sprachmodelle können Arbeit erleichtern und effektivere Prozesse ermöglichen. Bild-generierende KI-Tools können dabei unterstützen, Grafiken und Visualisierungen für Präsentationen, Webauftritte oder die sozialen Medien zu erstellen und den Nutzenden so Kapazitäten für andere Aufgaben freimachen.

Neueste Entwicklungen ermöglichen zudem die Nutzung von sogenannte Multimodalen KI-Systemen, die neben geschriebenem Text auch Audio und Video verarbeiten und generieren können. Solche Systeme können beispielsweise bei der Erstellung von Pressemitteilungen helfen, wo nicht nur Text, sondern auch Abbildungen und kurze Videos verwendet werden sollen.

Aufgrund der rasanten Entwicklung im Bereich (generativer) KI kann diese Liste nur eine Momentaufnahme der Möglichkeiten darstellen, die regelmäßig ergänzt werden muss. Damit geht einher, dass auch die damit verbundenen neuen Herausforderungen, Unsicherheiten und Risiken (siehe 4) gegenwärtig nur in groben Zügen zu umreißen sind. Die gesellschaftliche Einbettung von KI-Systemen befindet sich weitgehend noch in den Anfängen. Die Erfahrungen von Wissenschaft und Wissenschaftsmanagement mit den praktischen Möglichkeiten und Grenzen von KI-Systemen sind eine wichtige Ressource sozialer Reflexion und eine bedeutsame Instanz gesellschaftlichen Lernens hinsichtlich des zukünftigen Umgangs mit und der Gestaltung von solchen Systemen. Es ist das ausdrückliche Ziel der Helmholtz-Gemeinschaft, die Voraussetzungen dafür zu schaffen, dass sie und ihre Mitarbeitenden hierzu umfassend beitragen können.

4. Risiken und Handlungsempfehlungen zu deren Minimierung

KI-Systeme, insbesondere generative KI-Systeme, bieten viele Einsatzmöglichkeiten. Sie bergen jedoch auch Risiken und Herausforderungen. Einige davon basieren auf technischen Beschränkungen einzelner KI-Systeme, andere sind auf die (absichtliche oder unabsichtliche) missbräuchliche Verwendung von KI-Systemen zurückzuführen. Im Folgenden werden einzelne Risiken und Handlungsempfehlungen zu deren Minimierung aufgezeigt.

Grundsätzlich gilt, dass Nutzende von KI-Systemen immer selbst verantwortlich sind für

1. den Inhalt, den sie dem KI-System zur Verfügung stellen (Input), und

2. die Nutzung des Inhalts, den das KI-System herausgibt (Output).

KI-Systeme sind keine moralischen oder rechtlichen Akteure und können weder rechtlich noch moralisch für Erzeugnisse verantwortlich gemacht werden. Insbesondere können KI-Systeme keine Autorenschaft übernehmen und weder für Fehlinformationen, Datenschutzverletzungen, Urheberrechtsverletzungen, noch für andere rechtlich oder moralisch problematische Informationen zur Rechenschaft gezogen werden. Verantwortung setzt eine bewusste Entscheidungsfähigkeit und moralisches Urteilsvermögen voraus, die KI-Systeme nicht besitzen. Daher stehen die Nutzenden in der Verantwortung, KI-Systeme und deren Ergebnisse so zu verwenden, dass moralische Standards und rechtliche Vorschriften eingehalten werden. Bei Verstößen haften ggf. die Nutzenden.

4.1 Risiko: Datenschutzverletzung und ungewollte Weitergabe von Informationen

(Generative) KI-Systeme speichern und/oder nutzen häufig die von den Nutzenden eingegebenen Informationen, um z.B. die KI damit weiter zu trainieren. Auf diese Weise können die von Nutzenden eingegebenen Informationen sowohl zum Dienstleister als auch zu anderen, zukünftigen Nutzenden gelangen. Nutzende gehen bei bestimmten KI-Systemen somit das Risiko ein, sensible, vertrauliche oder persönliche Daten an andere weiterzugeben.

Dies kann unterschiedliche Konsequenzen nach sich ziehen:

- Die Datenschutz-Grundverordnung[2] (DSGVO) kann verletzt werden, welche dem Schutz von personenbezogenen Daten dient.
- Durch die Weitergabe von vertraulichen Geschäftsinformationen /-geheimnissen wird gegen das Geschäftsgeheimnisgesetz[3] verstoßen.
- Unveröffentlichte Forschungsergebnisse/-daten /-projekte und Manuskripte werden Dritten zugänglich gemacht, so dass diese von der Konkurrenz genutzt und ggf. veröffentlicht werden können.

[2] Nähere Informationen zur DSGVO finden Sie hier.
[3] Nähere Informationen zum Gesetz zum Schutz von Geschäftsgeheimnis finden Sie hier.

Handlungsempfehlung zur Risikominimierung

- Bei der Nutzung von (externen) KI-Systemen ist darauf zu achten, dass keine sensiblen Daten an die Dienstleister weitergegeben werden. Dazu gehören
  - Personenbezogene Daten (siehe DSGVO für Details)
  - Geschäftsgeheimnisse, vertrauliche/ strategische Geschäftsinformationen
  - Unveröffentlichte Forschungsergebnisse/-daten, wissenschaftliche Manuskripte, Forschungs-und Projektanträge
- Wenn dennoch KI-Systeme verwendet werden, ist darauf zu achten, dass die sensiblen Daten unkenntlich gemacht werden, sodass sie weder vom externen Dienstleister noch von anderen Nutzenden rekonstruiert oder genutzt werden können.
- Da zur sinnvollen Bewertung von wissenschaftlichen Manuskripten oder Forschungs- und Projektanträge alle Informationen des Dokuments nötig sind und daher sensible Daten nicht unkenntlich gemacht werden können, sind Bewertungen und Begutachtungen grundsätzlich ohne die Unterstützung von KI-Systemen vorzunehmen[4].

4.2 Risiko: Verletzung von Urheber- und sonstigen Schutzrechten

KI-Systeme können mit urheberrechtlich geschützten Daten oder unter Nutzung von sonstigem geistigen Eigentum trainiert worden sein, die teilweise oder vollständig in den generierten Ergebnissen wiedergegeben werden. Verwenden oder verbreiten Nutzende diese Ergebnisse (Bilder, Texte, Videos, Filme etc.) ungeprüft, kommt es zur Urheberrechtsverletzung[5] oder Verletzungen von anderen Rechten an geistigem Eigentum. In diesem Fall sind die Nutzenden verantwortlich und können ggf. rechtlich belangt werden, auch im Falle von unbeabsichtigter Verletzung von Urheberrechten und sonstigem geistigen Eigentum.

Handlungsempfehlung zur Risikominderung

- KI-generierte Inhalte sollten stets auf Marken und andere offensichtlich urheberrechtlich, markenrechtlich, oder anderweitig geschützte Elemente geprüft werden (z.B. Logos auf generierten Bildern), bevor sie weiterverarbeitet werden. Sollten geschützte Elemente in den Ergebnissen gefunden werden, sollte der betreffende Dienst und die damit generierten Inhalte nach Möglichkeit nicht genutzt werden.
- Es sollten bevorzugt etablierte Dienste genutzt werden, die auf Open-Source-Modellen basieren, ihre Trainingsdaten angeben und eine Überprüfung des Urheberrechtsstatus der Quelle ermöglichen.
- Bei der Erstellung von Inhalten mit hoher Sichtbarkeit/ Reichweite (z.B. Bilder für soziale Medien oder wissenschaftliche Publikationen), sollten Nutzende die Verwendung von KI-Tools stets sorgfältig gegen das Risiko einer unbeabsichtigten Urheberrechtsverletzung abwägen.

[4] Vgl. auch Stellungname der Deutschen Forschungsgemeinschaft (DFG)
[5] Nähere Informationen zum Urheberrecht finden Sie hier.

4.3 Risiko: Verbreitung von Falschinformationen und fehlende (wissenschaftliche) Informationsintegrität

Bei der Verwendung von KI-generierten Inhalten besteht das Risiko, dass Falschinformationen und Plagiate verbreitet werden. Diese Inhalte sind nicht mit den Grundzügen (wissenschaftlicher) Informationsintegrität vereinbar.

- **Falschinformationen:** KI-Systeme, insbesondere generative KI-Systeme, erzeugen ihre Ergebnisse, wie oben beschrieben, auf der Grundlage von Trainingsdaten, d.h. auf bereits bestehenden Texten, Daten, Bilder, Videos usw. Hierbei prüft das KI-System nicht, ob es sich bei den zum Training verwendeten Daten um korrekte Informationen handelt oder ob die Ergebnisse wahr oder sinnvoll sind. Es besteht folglich das Risiko, dass KI-generierte Inhalte falsche Informationen enthalten, welche bei deren Nutzung (unbeabsichtigt) verbreitet werden (können). Die Nutzenden stehen in der Verantwortung, den Output auf Richtigkeit zu prüfen, da sie bei Weiternutzung die Verantwortung dafür tragen. Die Offenlegungen und Dokumentation der Verwendung von KI-Systeme ist hier zentral.

- **„Halluzinationen“:** KI-Systeme können Elemente ihrer Trainingsdaten so zu neuen Inhalten zusammenfügen, dass der Output falsch oder sogar unsinnig ist. Bei „Halluzinationen“ handelt es sich also um eine spezielle Art der Falschinformation, welche vom KI-System selbst generiert werden und keine Grundlage in den zugrundeliegenden Daten oder der Realität haben. Nutzende dürfen Ergebnisse daher nicht ungeprüft übernehmen und sollten grundsätzlich eine kritische Einstellung gegenüber KI-generierten Erzeugnissen einnehmen.

- **Plagiate:** KI-generierte Inhalte können Plagiate enthalten, die direkt aus den Trainingsdaten stammen. Verwenden Nutzende diese Inhalte ungeprüft weiter, verbreiten sie Plagiate. Verwenden Nutzende KI-generierte Inhalte und geben dieses als ihr eigenes geistiges Eigentum aus, plagiieren sie, wenn die generierten Inhalte das geistige Eigentum anderer enthalten. Die Nutzenden tragen die Verantwortung für die genutzten Inhalte. Sie sind bei Weiterverarbeitung der generierten Inhalte die Autor:innen, nicht das verwendete KI-Tool.

Handlungsempfehlung zur Risikominderung

- Autorenschaft kann nur durch natürliche Personen übernommen werden. Diese sind es, die in ihrer Rolle der Autor:innen Verantwortung für Inhalte übernehmen und deren Qualität und Richtigkeit sicherstellen. Inhalte sind deshalb stets gründlich auf Konsistenz, Fehler und Plagiate zu prüfen.

- Die Anwendung von KI-Systemen ist offenzulegen sowie nachvollziehbar und transparent zu dokumentieren. Nutzende sollten also stets darüber informieren, wenn KI genutzt wurde (z.B. per Disclaimer, Fußnote, Hinweis, Zitation o.ä.). Diese Aufgabe zur Wahrung wissenschaftlicher Integrität ist ein Leitprinzip der guten wissenschaftlichen Praxis. In wissenschaftlichen Arbeiten müssen KI-Systeme im Detail zitiert werden, um z.B. Analysen von Daten nachvollziehbar und, soweit möglich, replizierbar zu machen.

- Open Science als leitendes Paradigma der wissenschaftlichen Arbeit in Helmholtz ist auch bei der Anwendung von KI-Systemen sicherzustellen. Über die Dokumentation hinaus sind Daten, Modelle und weiterführende Materialien nach dem Motto „as open as possible, as

closed as necessary“ entsprechend der Helmholtz Open Science Policy[6] offen und dokumentiert auf vertrauenswürdigen Repositorien zugänglich zu machen.

4.4 Risiko: Verzerrung / Vorurteile durch Trainingsdaten

Verzerrungen oder Vorurteile („Bias“), die in den Daten vorhanden sind, die zum Training eines KI-Systems verwendet wurden, können in den damit generierten Inhalten fortbestehen und sogar verstärkt werden. Dabei kann es sich um Vorurteile gegenüber ethnischen Gruppen, Religionsgemeinschaften usw. handeln, aber auch um Verzerrungen, die beeinflussen, welche Daten, Metadaten oder Quellen wie häufig und in welchem Umfang von KI-Systemen im Output verwendet werden. Durch die ungeprüfte Verwendung von KI-generierten Inhalten können Nutzende daher selbst diskriminierend agieren. Dies kann wiederum dazu führen, dass solche Verzerrungen und Vorurteile weiter verbreitet und somit langfristig verstärkt werden.

Handlungsempfehlung zur Risikominderung

- Bei der Verwendung von KI-generierten Inhalten sind Nutzende angehalten, diese sorgfältig auf Verzerrungen/ Vorurteile zu überprüfen. Dies gilt vor allem dann, wenn die KI-generierten Ergebnisse für Entscheidungen genutzt werden, die erhebliche Auswirkungen auf Dritte haben.

4.5 Risiko: Verstoß gegen KI-bezogene Regulierungen

KI wird vermehrt zu einer Massenware. Selbst mäßige Programmierkenntnisse reichen mittlerweile aus, um KI-Dienste einzurichten und bereitzustellen, und es ist zu erwarten, dass dies in Zukunft noch einfacher wird. Daher wächst das Risiko, unwissentlich Dienste zu verwenden und/oder zu entwickeln, die gesetzliche Regularien nicht oder nur unzulänglich umsetzen.

Handlungsempfehlung zur Risikominderung

- Der "EU AI Act“ ist das erste umfassende Gesetz, das einen Rechtsrahmen für den Einsatz von KI-Dienste schafft. Der EU AI Act[7] verfolgt einen risikobasierten Ansatz zur Regulierung von KI-Systemen und enthält, entsprechend der Einstufung eines KI-Systems, strenge Pflichten für Anbieter, Betreiber und Entwickler von KI-Systemen. Verstöße gegen diese Verpflichtungen können mit hohen Bußgeldern geahndet werden. Daher sollten Entwickler, Betreiber und Anbieter sich rechtlichen Rat einholen, bevor sie KI-Dienste anbieten oder betreiben.
- Nutzende sollten bei der Wahl von KI-Diensten darauf achten, dass diese die gesetzlichen Regularien vollumfänglich berücksichtigen.

[6]Siehe Helmholtz Open Science Policy
[7] Siehe EU AI Act

5. Good Practices bei der Nutzung & Entwicklung von KI-Systemen

Überprüfen von Inhalten: Alle KI-generierten Inhalte sollten sorgfältig von Nutzenden überprüft werden, um die oben aufgeführten Risiken zu minimieren. Nutzende sind für die von ihnen verwendeten/ verbreiteten Inhalte selbst verantwortlich und können bei gesetzlichen Verstößen ggf. haftbar gemacht werden. Dies gilt insbesondere dann, wenn die KI-generierten Ergebnisse der Entscheidungsfindung oder Bewertung dienen.

Transparenz und Nachvollziehbarkeit: Nutzende sollten stets transparent machen und offen dokumentieren, wann und in welchem Umfang KI genutzt wurde (z.B. per Disclaimer, Fußnote, Hinweis, Zitation o.ä.). Entwickler:innen von KI-Systemen sollten die grundlegende Funktionsweise sowie die fürs Training verwendeten Daten für Nutzende und Expert:innen ausreichend nachvollziehbar machen. Modelle und Daten sind im Sinne von Open Science auf vertrauenswürdigen Infrastrukturen, entsprechende dem Motto „as open as possible as closed as necessary" im Sinne der Helmholtz Open Science Policy zugänglich zu machen. Zudem sollten Nutzende stets darüber Informiert werden, wenn sie nicht mit einem Menschen, sondern einem KI-System interagieren.

Verantwortungsvolle (Nicht-)Nutzung: KI-Dienste sollten nicht verwendet werden, wenn die Gefahr besteht, dass sensible Daten (z.B. personenbezogene Daten, vertrauliche Geschäftsdaten, Geschäftsgeheimnisse, unveröffentlichte Forschungsergebnisse/-daten, Forschungsanträge usw.) weitergegeben werden könnten.

Verwendung bestimmter Dienste: Nutzende sollten nach Möglichkeit KI-Dienste verwenden, die transparent über ihre Datenquellen informieren. Aus datenschutzrechtlichen Gründen sollten Nutzende zudem KI-Systeme bevorzugen, die in Europa gehostet werden.

HELMHOLTZ

Spitzenforschung für
große Herausforderungen.

Impressum

Hermann von Helmholtz-Gemeinschaft Deutscher Forschungszentren e.V.
Geschäftsstelle Berlin
Anna-Louisa-Karsch-Straße 2, 10178 Berlin
www.helmholtz.de