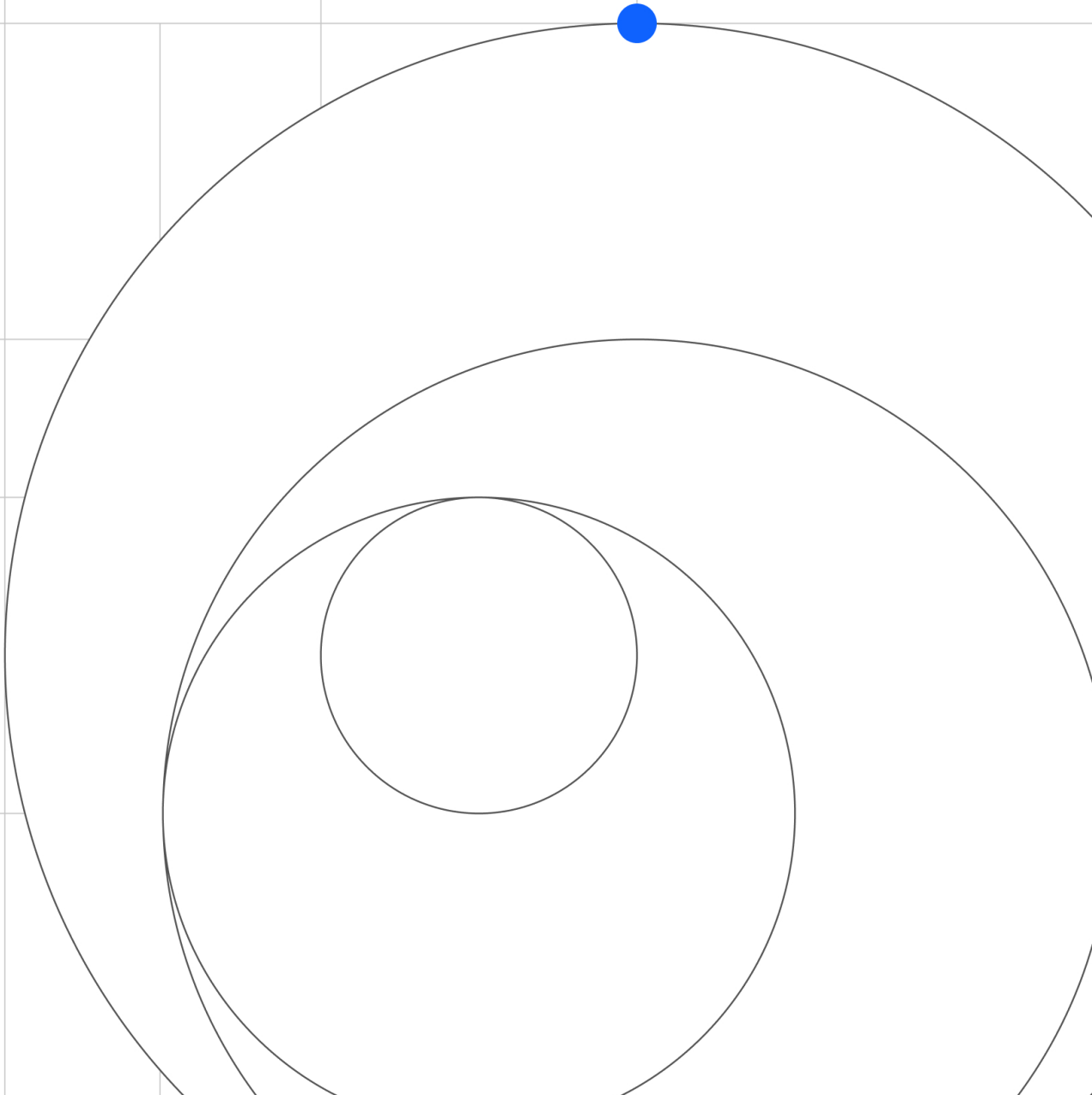
IBM AI Ethics Board

Basismodelle: Chancen, Risiken und Entschärfungen



Danksagung

Mit Dank an die leitenden Sponsoren des AI Ethics Board Workstreams, Christina Montgomery und Francesca Rossi, und die Beiträge der Workstream-Mitglieder Betsy Greytok, Bryan Bortnick, Catherine Quinlan, David Piorkowski, Eniko Rozsa, Heather Domin, Heather Gentile, Jamie VanDodick, Jill Maguire, John McBroom, Joshua New, Justin Weisz, Katherine Fick, Kevin Black, Kush Varshney, Manish Bhide, Manish Goyal, Melis Kiziltay, Michael Epstein, Michael Hind, Milena Pribic, Phaedra Boinodiris, Rogerio Abreu de Paula, Saishruthi Swaminathan und Suj Perepa.

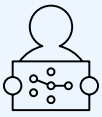
Inhaltsverzeichnis

Zusammenfassung

Das Vordringen von Basismodellen bietet Unternehmen aufregende neue Möglichkeiten, wirft aber auch neue und weitergehende Fragen in Bezug auf ihre ethische Gestaltung, Entwicklung, Bereitstellung und Nutzung auf. Laut einer kürzlich vom IBM Institute for Business Value durchgeführten Umfrage zum Thema generative KI in Bezug auf Vertrauensfragen, – insbesondere im Hinblick auf Investitionshemmnisse. Ihre Hauptsorgen sind Cybersicherheit (57 %), Datenschutz (51 %) und Genauigkeit (47 %). Viele Unternehmen haben diese Bedenken bereits vor der *Benutzeroptimierung* der generativen KI ernst genommen und ihre Absicht bekundet, in den nächsten drei Jahren mindestens 40 % mehr in die KI-Ethik zu investieren. Sich der Risiken bewusst zu werden und Wege zu finden, sie zu mindern, ist der erste entscheidende Schritt auf dem Weg zu vertrauenswürdigen KI-Systemen.

In diesem Dokument:

Untersuchung der Vorteile von Basismodellen, einschließlich ihrer Fähigkeit, anspruchsvolle Aufgaben zu erfüllen, das Potenzial zur Beschleunigung der Einführung von KI, die Fähigkeit zur Steigerung der Produktivität und die sich daraus ergebenen Kostenvorteile.



Erörterung der drei Risikokategorien, einschließlich der Risiken, die aus früheren Formen der KI bekannt sind, der bekannten Risiken, die durch Basismodelle verstärkt werden, und der neuen Risiken, die sich aus den generativen Fähigkeiten der Basismodelle ergeben.

Die Grundsätze, Säulen und die Governance, die die Grundlage der KI-Ethikinitiativen von IBM bilden, werden erläutert und Richtlinien zur Risikominderung vorgeschlagen.

Einführung

Mit der zunehmenden Verbreitung von KI liefern große und komplexe KI-Modelle vielversprechende Leistungsergebnisse und lösen einige der schwierigsten gesellschaftlichen Probleme. Die Erstellung großer Trainingsdatensätze und komplexer Modelle für jede KI-Anwendung kann für Unternehmen jedoch sehr aufwändig sein. Basismodelle bieten einen Weg, das Beste aus beiden Welten zu erreichen: leistungsfähige und technologisch ausgereifte Modelle zu erstellen und direkt wiederzuverwenden oder Tuning-Methoden anzuwenden, um eine Vielzahl von Anwendungsfällen zu realisieren, anstatt für jeden Anwendungsfall neue Modelle zu trainieren. IBM Research hat zum Beispiel Basismodelle für die visuelle Inspektion entwickelt. Diese Basismodelle lernen die allgemeine Beschaffenheit von Betonoberflächen und Landebahnen und können für bestimmte Anwendungsfälle wie die Risserkennung oder die Inspektion von Defekten mit weniger gekennzeichneten Daten weiter optimiert werden.

IBM definiert ein *Basismodell* als ein KI-Modell, das an eine Vielzahl nachgelagerter Aufgaben angepasst werden kann. Basismodelle sind in der Regel große generative Modelle, die mit Hilfe der Selbstüberwachung mit unmarkierten Daten trainiert werden. Als große Modelle können Basismodelle Milliarden von Parametern umfassen.

IBM ist ein hybrides Cloud- und KI-Unternehmen mit einer langen Tradition im verantwortungsvollen Umgang mit Daten und in der KI-Ethik. Mit der Stärke unserer Teams in den Bereichen Forschung, Produktentwicklung und Beratung sowie externen Partnern wie Hugging Face helfen wir unseren Kunden, die Leistungsfähigkeit von Basismodellen zu nutzen und vertrauenswürdige KI in jedem Unternehmen zu etablieren. IBM investiert auch weiterhin in den Aufbau neuer Plattformen, wie die KI- und Datenplattform IBM watsonx™, und Technologien für das Design und die Entwicklung von KI-Modellen, die sich auf eine protokollierbare und vertrauenswürdige Weise verhalten.

Dieses Dokument beschreibt die Position von IBM zur Ethik von Basismodellen. Dies ist die erste Version. Zukünftige Versionen werden verschiedene Aspekte des ethischen Ansatzes des IBM Basismodells erweitern. Wir hoffen, dass dieses Dokument allen Beteiligten helfen wird, das Basismodell verantwortungsvoll zu entwickeln, anzuwenden und zu nutzen.

Vorteile von Basismodellen

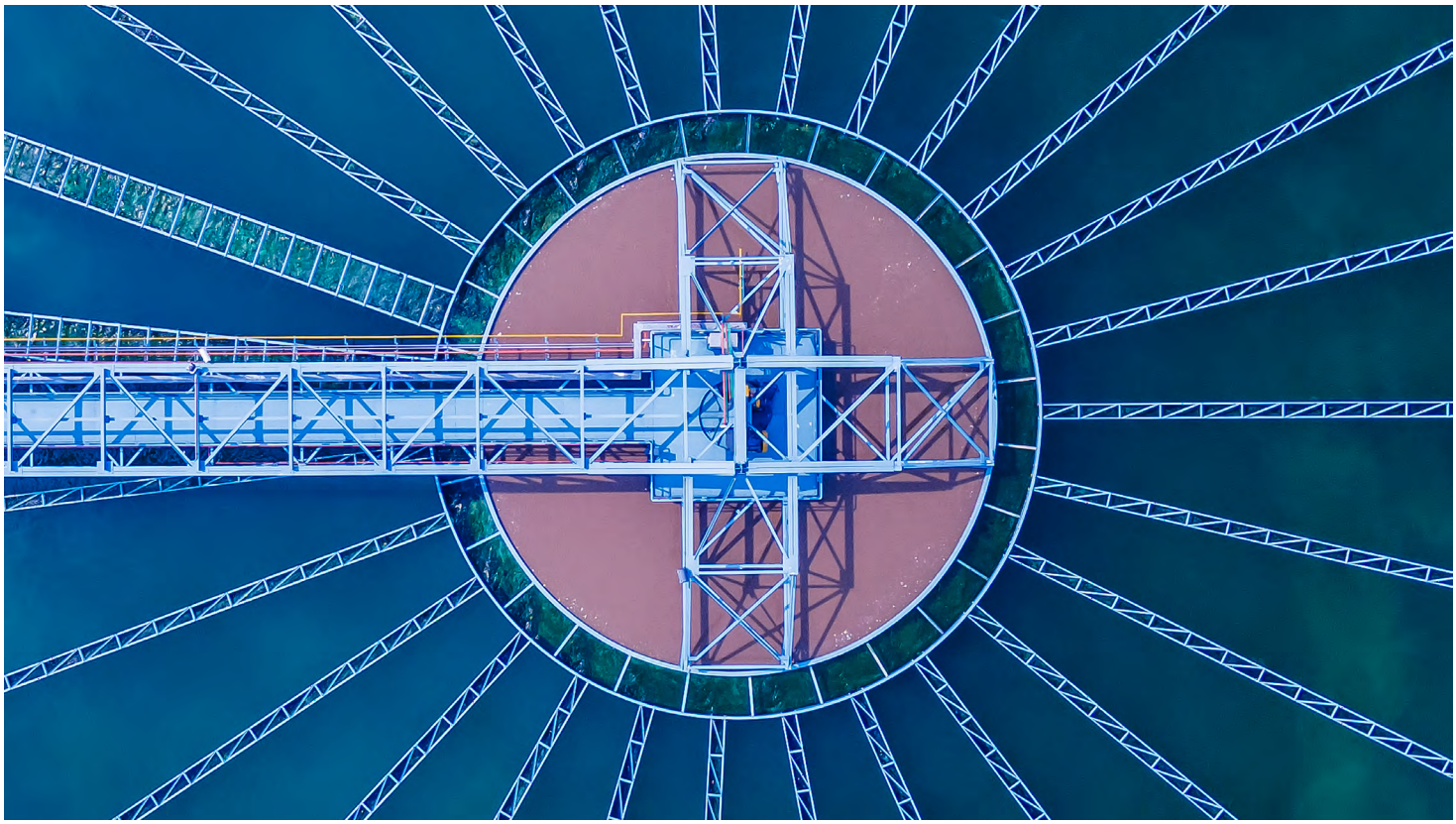
Basismodelle können den Entwicklungsprozess von KI-Systemen erheblich verbessern und dazu beitragen, dass KI in Unternehmen von der Explorationsphase in die Akzeptanzphase übergeht. Zu den Vorteilen zählen:

Ausführen komplexer Aufgaben
Basismodelle zeigen eine signifikante Leistungssteigerung bei der Lösung schwieriger und komplexer Probleme. So soll beispielsweise das georäumliche Basismodell der Zusammenarbeit von IBM und NASA Satellitendaten der NASA in Karten von Naturkatastrophen wie Überschwemmungen und anderen Landschaftsveränderungen umwandeln. Das Modell könnte auch genutzt werden, um die Vergangenheit unseres Planeten zu erforschen, Risiken für Ernten, Unternehmen oder Infrastrukturen durch Unwetter abzuschätzen, Strategien zur Anpassung an den Klimawandel zu entwickeln und die Agrarindustrie zu unterstützen. Eine Vorschau des Modells wird IBM-Kunden über die IBM Environmental Intelligence Suite zur Verfügung gestellt.

Ein weiteres Beispiel ist der MoLFormer-XL von IBM. Hierbei handelt es sich um ein Basismodell, das die Struktur von Molekülen aus einfachen Darstellungen ableitet und das Erlernen verschiedener nachgelagerter Aufgaben erleichtert, z. B. die Vorhersage der physikalischen und quantitativen Eigenschaften eines Moleküls, die Identifizierung ähnlicher Moleküle, das Screening bereits zugelassener Moleküle für neue Anwendungsfälle und die Entdeckung neuer Moleküle. Moderna und IBM untersuchen gemeinsam, wie MoLFormer dazu beitragen kann, die Eigenschaften von Molekülen vorherzusagen und die Eigenschaften potenzieller mRNA-Medikamente zu verstehen.

Produktivitätssteigerung
Die generative Natur der Basismodelle erweitert die Anzahl der Bereiche, in denen KI in einem Unternehmen eingesetzt werden kann, um die Produktivität zu steigern, indem Routine- und mühsame Aufgaben automatisiert werden und den Benutzern mehr Zeit für kreative und innovative Arbeit bleibt. Beispielsweise ermöglicht IBM Watson Code Assistent, der auf Basismodellen basiert, Entwicklern aller Erfahrungsstufen das Schreiben von Code mithilfe von KI-generierten Empfehlungen.

Kürzere Zeit bis zur Wertschöpfung
Basismodelle werden in der Regel mit unmarkierten Daten trainiert, die in größerer Menge zur Verfügung stehen als markierte Daten. Einmal trainierte Basismodelle können entweder direkt oder nach einer Anpassung für nachgelagerte Anwendungen verwendet werden, wobei nur eine geringe Menge an speziell markierten Daten benötigt wird, was die Zeit bis zur Wertschöpfung verkürzen kann.



Verwendung verschiedener Datenmodalitäten
Basismodelle können mit verschiedenen Datenmodalitäten trainiert werden, z. B. mit natürlicher Sprache, Text, Bild und Audio. Sie können auch auf Aufgaben angewendet werden, die verschiedene Datentypen erfordern, z. B. Zeitreihendaten, Geodaten, Tabellendaten, halbstrukturierte Daten und gemischte Datenmodalitäten wie Text in Kombination mit Bildern.

Amortisierte Kosten
Obwohl die anfänglichen Kosten für das Training eines Basismodells deutlich höher sind als die eines herkömmlichen KI-Modells, sind die zusätzlichen Kosten für die Anwendung auf eine neue Aufgabe wesentlich geringer. Die Verwendung vortrainierter Basismodelle könnte dazu beitragen, dass Unternehmen keine erheblichen Investitionen tätigen müssen, um Basismodelle zu trainieren und mit ihren neuen Fähigkeiten zu experimentieren. Für ein Unternehmen sind die Zuverlässigkeit der Modelle, die Energieeffizienz, die Leistungsfähigkeit, die Übertragbarkeit und die Fähigkeit, Unternehmensdaten effizient und sicher zu nutzen, von größter Bedeutung.

IBM ermöglicht es Unternehmen, durch die Nutzung der besten Innovationen aus der offenen, globalen KI-Community, durch effizientes Arbeiten in hybriden IT-Umgebungen, durch Risikominimierung und durch strenge Kontrolle der KI den Wert von Basismodellen für ihr Unternehmen zu schaffen und zu nutzen.

Risiken von Basismodellen

Wie alle schnelllebigen Technologien bergen Foundation Models Vorteile und Risiken. Bei einigen handelt es sich um rechtliche Risiken, z. B. Beschränkungen bei der Übermittlung oder Nutzung von Daten, die nach geltendem und sich entwickelndem Recht sorgfältig geprüft werden müssen. Andere Risiken sind ethischer Natur und müssen sorgfältig abgewogen werden, wenn die Technologie positive Auswirkungen haben soll. Generell werfen KI-Risiken soziotechnische Fragen auf und sollten mit soziotechnischen Methoden adressiert werden. Dazu gehören Software-Tools, Risikobewertungsprozesse, ethische Frameworks für KI, Governance-Mechanismen, Konsultationen mit verschiedenen Interessengruppen, Normen und Vorschriften. Die Risiken werden unter Berücksichtigung der folgenden 3 Kategorien aufgelistet:

1. **Traditionell.** Bekannte Risiken aus vorherigen oder früheren Formen von KI-Systemen
2. **Verstärkt.** Die Risiken sind bekannt, werden aber durch die inhärenten Eigenschaften der Basismodelle, insbesondere die zugrunde liegende Fähigkeit zur Generierung, noch verstärkt
3. **Neu.** Aufkommende Risiken, die sich aus den Basismodellen und der ihnen zugrunde liegenden Fähigkeit zur Generierung ergeben

Wir strukturieren die Liste der Risiken auch danach, ob sie hauptsächlich mit den Inhalten verbunden sind, die dem Foundation Model zur Verfügung gestellt werden – dem Input – oder mit den Inhalten, die es erzeugt – dem Output – oder ob sie mit zusätzlichen Herausforderungen verbunden sind.

1. Risiken im Zusammenhang mit der Eingabe

Trainings- und Optimierungsphase

Gruppe	Risiko	Warum ist dies ein Grund zur Sorge?	Indikator
Gerechtigkeit	Daten-Bias: Historische, repräsentative und gesellschaftliche Verzerrungen in den Daten, die zum Training und zur Feinabstimmung des Modells verwendet werden.	Das Trainieren eines KI-Systems auf Daten mit Bias, wie z. B. historischem oder repräsentativem Bias, könnte zu voreingenommenen oder verzerrten Outputs führen, die möglicherweise bestimmte Gruppen oder Einzelpersonen unfair darstellen oder anderweitig diskriminieren. Zusätzlich zu den negativen Auswirkungen auf die Gesellschaft könnten Unternehmen aufgrund von verzerrten Modellergebnissen rechtliche Konsequenzen, Betriebsunterbrechungen oder Reputationsschäden erleiden.	Verstärkt
Zuverlässigkeit	Data Poisoning: eine Art feindlicher Angriff, bei dem ein Angreifer oder böswilliger Insider absichtlich beschädigte, falsche, irreführende oder fehlerhafte Proben in den Trainings- oder Feinabstimmungs-Datensatz einbringt.	Data Poisoning erhöht die Empfänglichkeit eines Modells für ein bösartiges Datenmuster und bringt den vom Angreifer gewünschten Output hervor. Dies kann ein Sicherheitsrisiko darstellen, da Angreifer das Modellverhalten zu ihrem eigenen Vorteil erzwingen können. Neben unbeabsichtigten und potenziell böswilligen Ergebnissen kann eine Modellfehlanpassung aufgrund von Data Poisoning dazu führen, dass Unternehmen mit rechtlichen Konsequenzen, Betriebsunterbrechungen oder Reputationsschäden konfrontiert werden.	Traditionell
Werteausrichtung	Datenpflege: Wenn Trainings- oder Abstimmungsdaten unsachgemäß gesammelt oder aufbereitet wurden.	Eine unsachgemäße Datenpflege kann sich negativ auf das Trainieren eines Modells auswirken und zu einem Modell führen, das sich nicht entsprechend den beabsichtigten Werten verhält. Beispiele für eine unsachgemäße Datenpflege könnten Beschriftungs- oder Anmerkungsfehler in den für das Training oder die Abstimmung des Modells verwendeten Daten sein. Die Korrektur von Problemen, nachdem das Modell trainiert und bereitgestellt wurde, reicht möglicherweise nicht aus, um ein korrektes Verhalten sicherzustellen. Unangemessenes Modellverhalten kann dazu führen, dass Unternehmen mit rechtlichen Konsequenzen, Betriebsunterbrechungen oder Reputationsschäden konfrontiert werden.	Verstärkt
	Downstream-basiertes Retraining: Verwendung unerwünschter (ungenauer, unangemessener, benutzereigener Inhalte usw.) Outputs von nachgelagerten Anwendungen zu Retraining-Zwecken.	Wiederverwendung von Downstream-Ausgaben zum erneuten Trainieren eines Modells ohne eine angemessene manuelle Überprüfung erhöht die Wahrscheinlichkeit, dass unerwünschte Outputs in die Trainings- oder Abstimmungsdaten des Modells einfließen, wodurch möglicherweise noch mehr unerwünschter Output erzeugt wird. Unangemessenes Modellverhalten kann dazu führen, dass Unternehmen mit rechtlichen Konsequenzen oder Reputationsschäden rechnen müssen. Die Nichteinhaltung der Gesetze zur Datenübermittlung kann zu Geldstrafen und anderen rechtlichen Konsequenzen führen.	Neu
Datengesetze	Datenübertragung: Gesetze und andere Einschränkungen können die Übertragung von Daten einschränken oder verbieten.	Einschränkungen bei der Datenübertragung können die Verfügbarkeit der für das Training eines KI-Modells benötigten Daten beeinträchtigen und zu schlecht dargestellten Daten führen. Zusätzlich zu den Auswirkungen auf die Datenverfügbarkeit kann die Nichteinhaltung von Gesetzen und Vorschriften zur Datenübertragung zu Geldstrafen und anderen rechtlichen Konsequenzen führen.	Traditionell
	Datennutzung: Gesetze und andere Einschränkungen können die Verwendung einiger Daten für bestimmte KI-Anwendungsfälle einschränken oder verbieten.	Die Nichteinhaltung von Gesetzen und Vorschriften zur Datennutzung kann Geldstrafen und andere rechtliche Konsequenzen nach sich ziehen.	Traditionell
	Datenakquisition: Gesetze und andere Vorschriften können die Erfassung bestimmter Arten von Daten für bestimmte KI-Anwendungsfälle einschränken.	Die Nichteinhaltung von Gesetzen und Vorschriften zur Datenakquisition kann zu Geldstrafen und anderen rechtlichen Konsequenzen führen.	Verstärkt

Gruppe	Risiko	Warum ist dies ein Grund zur Sorge?	Indikator
Geistiges Eigentum	Rechte zur Datennutzung: Nutzungsbedingungen, Urheberrechtsgesetze, die Einhaltung von Lizenzen oder andere Fragen des geistigen Eigentums können die Möglichkeit einschränken, bestimmte Daten für die Erstellung von Modellen zu verwenden.	Die Gesetze und Vorschriften für die Verwendung von Daten zum Trainieren von KI sind ungeklärt und können von Land zu Land variieren, was zu Herausforderungen bei der Entwicklung von Modellen führt. Wenn die Datennutzung gegen Regeln oder Beschränkungen verstößt, müssen Unternehmen mit Geldstrafen, Rufschädigung, Betriebsunterbrechungen und anderen rechtlichen Konsequenzen rechnen.	Verstärkt
Transparenz	Datentransparenz: Die Herausforderung besteht in der Dokumentation, wie die Daten eines Modells erfasst, kuratiert und zum Trainieren eines Modells verwendet wurden.	Datentransparenz ist für die Einhaltung von Gesetzen und die KI-Ethik wichtig. Fehlende Informationen schränken die Möglichkeit ein, die mit den Daten verbundenen Risiken zu bewerten. Das Fehlen standardisierter Anforderungen könnte die Transparenz beeinträchtigen, da Unternehmen ihre Betriebsgeheimnisse schützen und versuchen, andere am Kopieren ihrer Geschäftsmodelle zu hindern.	Verstärkt
	Datenherkunft: Die Herausforderung besteht in der Standardisierung und Einführung von Methoden zur Überprüfung der Herkunft von Daten.	Nicht alle Datenquellen sind vertrauenswürdig. Die Daten könnten auf unethische Weise erhoben, manipuliert oder gefälscht worden sein. Dementsprechend kann die Verwendung unzuverlässiger Daten zu unerwünschten Verhaltensweisen im Modell führen. Unternehmen müssen mit Geldstrafen, Rufschädigung, Betriebsunterbrechungen und anderen rechtlichen Konsequenzen rechnen.	Verstärkt
Datenschutz	Personenbezogene Informationen in Daten: Einschluss oder Vorhandensein personenbezogener Informationen (Personal Identifiable Information, PII) und sensiblen persönlichen Informationen (Sensitive Personal Information, SPI) in den Daten, die für das Training oder die Feinabstimmung des Modells verwendet werden.	Wenn das Modell nicht ordnungsgemäß entwickelt wurde, um sensible Daten zu schützen, könnte es im generierten Output persönliche Informationen preisgeben. Darüber hinaus müssen personenbezogene oder sensible Daten gemäß den Datenschutzgesetzen und -vorschriften geprüft und gehandhabt werden. Unternehmen müssen bei Verstößen mit Geldbußen, Rufschädigung, Betriebsunterbrechungen und anderen rechtlichen Konsequenzen rechnen.	Traditionell
	Re-Identifizierung: Selbst wenn personenbezogene Informationen (PII) und sensible persönliche Informationen (SPI) aus den Daten entfernt wurden, kann es immer noch möglich sein, Personen aufgrund anderer in den Daten vorhandener Funktionen zu identifizieren.	Daten, die persönliche oder sensible Informationen offenlegen können, müssen im Hinblick auf Datenschutzgesetze und -vorschriften überprüft werden, da Unternehmen bei Verstößen mit Geldstrafen, Rufschädigung, Betriebsunterbrechungen und anderen rechtlichen Konsequenzen rechnen müssen.	Traditionell
	Datenschutzrechte: Herausforderungen in Bezug auf die Möglichkeit, den Betroffenen Rechte zu gewähren, wie beispielsweise das Recht auf Widerspruch, das Recht auf Zugang und das Recht auf Vergessenwerden.	Die Identifizierung oder missbräuchliche Verwendung von Daten könnte zur Verletzung von Datenschutzgesetzen führen. Eine unsachgemäße Verwendung oder eine Aufforderung zur Datenlöschung könnte Unternehmen dazu zwingen, das Modell erneut zu trainieren, was teuer ist. Darüber hinaus können Unternehmen mit Geldstrafen, Rufschädigung, Betriebsunterbrechungen und anderen rechtlichen Konsequenzen rechnen, wenn sie die Datenschutzbestimmungen nicht einhalten.	Verstärkt
	Einverständniserklärung: Daten, die für das Training von KI-Modellen ohne die informierte Zustimmung des Eigentümers gesammelt werden, selbst wenn dies gesetzlich erlaubt ist.	Unter bestimmten Umständen kann es unethisch sein, Daten ohne die Zustimmung der Person zu sammeln und zu verwenden. Eine solche Verwendung birgt auch mögliche Reputationsrisiken.	Traditionell

Inferenz Phase

Gruppe	Risiko	Warum ist dies ein Grund zur Sorge?	Indikator
Datenschutz	Personenbezogene Informationen im Prompt: Offenlegung personenbezogener oder sensibler personenbezogener Informationen als Teil einer Aufforderung an das Modell.	Die Prompt-Daten können gespeichert oder später für andere Zwecke wie die Modellevaluation und das Retraining verwendet werden. Diese Arten von Daten müssen im Hinblick auf Datenschutzgesetze und -vorschriften überprüft werden. Ohne ordnungsgemäße Datenspeicherung und -nutzung drohen Unternehmen Geldstrafen, Rufschädigung, Betriebsunterbrechungen und andere rechtliche Konsequenzen.	Neu
Geistiges Eigentum	IP-Informationen im Prompt: Offenlegung von Informationen zu Urheberrechten oder anderen IP-Informationen als Teil der Systemanfrage, die an das Modell gesendet wurde.	Die Prompt-Daten können gespeichert oder später für andere Zwecke wie die Modellevaluation und das Retraining verwendet werden. Diese Arten von Daten müssen im Hinblick auf Gesetze und Vorschriften zum geistigen Eigentum überprüft werden. Ohne ordnungsgemäße Datenspeicherung und -nutzung müssen Unternehmen mit Geldstrafen, Rufschädigung, Betriebsunterbrechungen und anderen rechtlichen Konsequenzen rechnen.	Neu
	Vertrauliche Daten im Prompt: Aufnahme vertraulicher Daten als Teil des Prompts, der an das Modell gesendet wird.	Wenn das Modell nicht ordnungsgemäß entwickelt wurde, um vertrauliche Daten zu schützen, könnte es vertrauliche Informationen oder geistiges Eigentum im generierten Output preisgeben. Außerdem können vertrauliche Informationen der Endbenutzer unbeabsichtigt erfasst und gespeichert werden.	Neu
Zuverlässigkeit	Umgehungsangriff: Versuch, ein Modell dazu zu bringen, falsche Outputs auszugeben, indem die an das trainierte Modell gesendeten Daten verfälscht werden.	Umgehungsangriffe verändern das Verhalten des Modells, meist zum Vorteil des Angreifers. Wenn die Output-Ergebnisse nicht ordnungsgemäß verbucht werden, müssen Unternehmen mit Geldstrafen, Rufschädigung, Betriebsunterbrechungen und anderen rechtlichen Konsequenzen rechnen.	Verstärkt
	Prompt-basierte Angriffe: Angriffe wie Prompt Injection (Versuch, ein Modell zu zwingen, einen unerwarteten Output zu erzeugen), Prompt Leaking (Versuch, den Systemprompt eines Modells zu extrahieren), Jailbreaking (Versuch, die im Modell eingerichteten Sicherheitsbarrieren zu durchbrechen) und Prompt Priming (Versuch, ein Modell zu zwingen, einen Output zu erzeugen, der auf den Prompt passt).	Abhängig von den offenbarten Inhalten können Unternehmen mit Geldstrafen, Rufschädigung, Betriebsunterbrechungen und anderen rechtlichen Konsequenzen rechnen.	Neu

2. Risiken im Zusammenhang mit der Ausgabe

Gruppe	Risiko	Warum ist dies ein Grund zur Sorge?	Indikator
Gerechtigkeit	Output-Bias: Bei der Erstellung von Inhalten kann es zu einer unfairen Vertretung bestimmter Gruppen oder Einzelpersonen kommen.	Bias kann den Nutzern der KI-Modelle schaden und bestehende diskriminierende Verhaltensweisen verstärken. Unternehmen müssen mit Rufschädigung, Betriebsunterbrechungen und anderen Konsequenzen rechnen.	Neu
	Entscheidungs-Bias: Wenn eine Gruppe aufgrund der Auswirkungen von Entscheidungen, die von Menschen auf der Grundlage des Modell-Outputs getroffen werden, gegenüber einer anderen Gruppe ungerechtfertigt bevorzugt wird.	Bias kann Personen schaden, die von den Entscheidungen des Modells betroffen sind. Unternehmen müssen mit Geldstrafen, Rufschädigung, Betriebsunterbrechungen und anderen rechtlichen Konsequenzen rechnen.	Traditionell
Geistiges Eigentum	Verletzung des Urheberrechts: Wenn ein Modell Inhalte generiert, die einem bestehenden Werk, das urheberrechtlich geschützt ist oder unter eine Open-Source-Lizenzvereinbarung fällt, zu ähnlich oder identisch sind.	Die Gesetze und Vorschriften für die Verwendung von Inhalten, die anderen urheberrechtlich geschützten Daten gleich oder sehr ähnlich sind, sind weitgehend ungeklärt und können von Land zu Land variieren, was die Festlegung und Umsetzung der Compliance erschwert. Unternehmen müssen mit Geldstrafen, Rufschädigung, Betriebsunterbrechungen und anderen rechtlichen Konsequenzen rechnen.	Neu
Werteausrichtung	Halluzination: Erzeugung von Inhalten, die sachlich unrichtig oder unwahr sind.	Falsche Outputs können Benutzer in die Irre führen und in nachgelagerte Artefakte einfließen, wodurch Fehlinformationen weiter verbreitet werden. Dies kann sowohl den Besitzern als auch den Nutzern der KI-Modelle schaden. Zudem müssen Unternehmen mit Geldstrafen, Rufschädigung, Betriebsunterbrechungen und anderen rechtlichen Konsequenzen rechnen.	Neu
	Toxischer Output: Wenn das Modell hasserfüllte, beleidigende und profane (Hateful, Abusive, Profane – HAP) oder obszöne Inhalte produziert.	Hasserfüllte, beleidigende und profane (HAP) oder obszöne Inhalte können Menschen, die mit dem Modell interagieren, negativ beeinflussen und schädigen. Zudem müssen Unternehmen mit Geldstrafen, Rufschädigung, Betriebsunterbrechungen und anderen rechtlichen Konsequenzen rechnen.	Neu
	Gefährlicher Ratgeber: Wenn ein Modell berät, ohne über genügend Informationen zu verfügen, was zu möglichen Gefahren bei Befolgung dieses Ratschlags führt.	Eine Person könnte auf einen unvollständigen Ratschlag hin handeln oder sich Sorgen über eine Situation machen, die nicht auf sie zutrifft, weil der generierte Inhalt zu allgemein gehalten ist.	Neu
Missbrauch	Verbreiten von Desinformation: Die Verwendung eines Modells zur Erstellung irreführender oder falscher Informationen, um eine Zielgruppe zu täuschen oder zu beeinflussen.	Die Verbreitung von Desinformationen kann die Fähigkeit eines Menschen beeinträchtigen, fundierte Entscheidungen zu treffen. Unternehmen müssen mit Geldstrafen, Rufschädigung, Betriebsunterbrechungen und anderen rechtlichen Konsequenzen rechnen.	Neu
	Toxizität: Verwendung eines Modells, das hasserfüllte, beleidigende und profane (HAP) oder obszöne Inhalte erzeugt.	Toxische Inhalte können sich negativ auf das Wohlbefinden der Empfänger auswirken. Unternehmen müssen mit Geldstrafen, Rufschädigung, Betriebsunterbrechungen und anderen rechtlichen Konsequenzen rechnen.	Neu
	Nicht-einvernehmliche Nutzung: Die Verwendung eines Modells zur Imitation von Personen durch Video (Deepfakes), Bilder, Audio oder andere Modalitäten ohne deren Zustimmung.	Deepfakes können Desinformationen über eine Person verbreiten, was sich möglicherweise negativ auf den Ruf der Person auswirkt. Unternehmen müssen mit Geldstrafen, Rufschädigung, Betriebsunterbrechungen und anderen rechtlichen Konsequenzen rechnen.	Verstärkt

Gruppe	Risiko	Warum ist dies ein Grund zur Sorge?	Indikator
	Gefährliche Anwendung: Die Verwendung eines Modells mit der alleinigen Absicht, Menschen zu schaden.	Unternehmen müssen mit Geldstrafen, Rufschädigung, Betriebsunterbrechungen und anderen rechtlichen Konsequenzen rechnen.	Neu
	Geheimhaltung: Kein Hinweis darauf, dass der Inhalt von einem KI-Modell erzeugt wurde.	Das Verschweigen der von der KI erstellten Inhalte kann als trügerisch angesehen werden und zu einem Vertrauensverlust führen. Vorsätzliche Täuschung kann zu einer Verringerung der menschlichen Handlungsfähigkeit, Geldstrafen, Rufschädigung und anderen rechtlichen Konsequenzen führen.	Neu
	Unsachgemäße Verwendung: Verwendung eines Modells für einen Zweck, für den das Modell nicht konzipiert wurde.	Die Wiederverwendung eines Modells ohne Kenntnis der ursprünglichen Daten, der Designabsicht und der Ziele kann zu unerwartetem und unerwünschtem Modellverhalten führen.	Verstärkt
Schädigende Codegener-ierung	Erzeugung von schädlichem Code: Modelle können Code generieren, der bei seiner Ausführung Schaden anrichtet oder andere Systeme ungewollt beeinträchtigt.	Die Ausführung von schädlichem Code kann Schwachstellen in IT-Systemen öffnen. Unternehmen müssen mit Geldstrafen, Rufschädigung, Betriebsunterbrechungen und anderen rechtlichen Konsequenzen rechnen.	Neu
Unangebrachtes Vertrauen	Zu großes/zu geringes Vertrauen: Wenn eine Person zu wenig oder zu viel Vertrauen in die Anleitung eines KI-Modells setzt.	Bei Aufgaben, bei denen Menschen Entscheidungen auf der Grundlage von KI-basierten Vorschlägen treffen, kann ein zu großes oder zu geringes Vertrauen in das KI-System zu einer schlechten Entscheidungsfindung führen. Dies wiederum kann negative Folgen nach sich ziehen, die mit der Bedeutung der Entscheidung zunehmen. Schlechte Entscheidungen können Menschen schaden und zu finanziellem Schaden, Rufschädigung, Betriebsunterbrechungen und anderen rechtlichen Konsequenzen für Unternehmen führen.	Verstärkt
Datenschutz	Offenlegung personenbezogener Informationen: Wenn personenbezogene Daten (PII) oder sensible persönliche Daten (SPI) in den Trainingsdaten, den Feinabstimmungsdaten oder als Teil der Eingabeaufforderung verwendet werden, geben die Modelle diese Daten möglicherweise im generierten Output preis.	Die Weitergabe von persönlichen Informationen beeinträchtigt die Rechte der Menschen und erhöht die Angreifbarkeit. Außerdem müssen die Output-Daten im Hinblick auf Datenschutzgesetze und -vorschriften überprüft werden, da Unternehmen mit Geldstrafen, Rufschädigung, Betriebsunterbrechungen und anderen rechtlichen Konsequenzen rechnen müssen, wenn sie gegen Datenschutz- oder Nutzungsgesetze verstoßen.	Neu
Erklärbarkeit	Unerklärlicher Output: Herausforderungen bei der Erklärung, warum der Modell-Output generiert wurde.	Foundation Models basieren auf komplexen Deep-Learning-Architekturen, was es schwer macht, ihre Outputs zu erläutern. Ohne klare Erklärungen für den Modell-Output ist es für Benutzer, Modellvalidierer und Auditoren schwierig, das Modell zu verstehen und ihm zu vertrauen. Mangelnde Transparenz kann in stark regulierten Bereichen rechtliche Konsequenzen nach sich ziehen. Falsche Erklärungen können zu übermäßigem Vertrauen führen.	Verstärkt
Rückverfolg-barkeit	Unzuverlässige Zuweisung von Quellen: Es ist schwierig festzustellen, aus welchen Trainings- oder Feinabstimmungsdaten das Modell einen Teil oder den gesamten Output generiert hat.	Wenn die Quelle oder Herkunft eines Outputs nicht zurückverfolgt werden kann, fällt es Benutzern, Modell-Validierern und Prüfern schwer, das Modell zu verstehen und ihm zu vertrauen.	Neu

3. Herausforderungen

Gruppe	Risiko	Warum ist dies ein Grund zur Sorge?	Indikator
Governance	Modell-Transparenz: Mangelnde Modell-Transparenz oder unzureichende Dokumentation des Modellentwicklungsprozesses macht es schwierig zu verstehen, wie und warum ein Modell erstellt wurde und wer es erstellt hat, was die Möglichkeit eines unbeabsichtigten Missbrauchs des Modells erhöht.	Transparenz ist für die Einhaltung von Gesetzen, KI-Ethik und die angemessene Nutzung von Modellen wichtig. Fehlende Informationen können die Risikobewertung, die Änderung des Modells oder seine Wiederverwendung erschweren. Das Wissen darüber, wer ein Modell entwickelt hat, kann ebenfalls ein wichtiger Faktor für die Entscheidung sein, ob man dem Modell vertraut.	Traditionell
	Verantwortlichkeit: Der Entwicklungsprozess von Foundation Models ist komplex und umfasst eine Vielzahl von Daten, Prozessen und Rollen. Funktioniert der Output des Modells nicht wie erwartet, kann es schwierig sein, die Grundursache zu ermitteln und Verantwortlichkeiten zuzuweisen.	Ohne die ordnungsgemäße Dokumentation von Entscheidungen und die Zuweisung von Verantwortlichkeiten ist es unter Umständen nicht möglich, die Haftung für unerwartetes Verhalten oder Missbrauch zu bestimmen.	Verstärkt
Einhaltung gesetzlicher Vorschriften	Rechtliche Verantwortlichkeit: Bestimmen Sie, wer für das Foundation Model verantwortlich ist.	Wenn nicht klar ist, wer die Verantwortung für die Entwicklung des Modells trägt, besteht die Gefahr, dass Aufsichtsbehörden und andere Personen Bedenken gegen das Modell haben. Es ist nicht klar, wer für Probleme mit dem Modell haftet bzw. wer verantwortlich ist – bzw. sein sollte – oder wer Fragen zu dem Modell beantworten kann. Nutzer von Modellen ohne eindeutige Eigentumsverhältnisse werden möglicherweise Probleme mit der Einhaltung künftiger KI-Vorschriften haben.	Neu
	Eigentumsrechte an generierten Inhalten: Bestimmung der Eigentumsrechte an KI-generierten Inhalten.	Gesetze und Vorschriften, die sich auf die Eigentumsrechte an KI-generierten Inhalten beziehen, sind weitgehend ungeklärt und können von Land zu Land variieren. Unternehmen müssen mit Geldstrafen, Reputationsrisiken, Betriebsunterbrechungen und anderen rechtlichen Konsequenzen rechnen.	Neu
	IP für generierte Inhalte: Rechtsunsicherheit über geistige Eigentumsrechte im Zusammenhang mit generierten Inhalten.	Die Gesetze und Vorschriften zur Bestimmung der Urheberrechtsfähigkeit und Patentierbarkeit von KI-generierten Inhalten sind weitgehend ungeklärt und können von Land zu Land variieren. Unternehmen müssen mit Geldstrafen, Reputationsrisiken, Betriebsunterbrechungen und anderen rechtlichen Konsequenzen rechnen, wenn die generierten Inhalte durch geistige Eigentumsrechte geschützt sind.	Neu
	Quellennachweis: Bestimmung der Herkunft der generierten Inhalte.	Wenn das Modell einen Output erzeugt, der mit den Daten identisch ist, die zum Trainieren des Modells verwendet wurden, sollte es die Herkunft dieses Outputs angeben. Andernfalls können die Unternehmen, die das Modell bereitstellen oder verwenden, einem rechtlichen Risiko ausgesetzt sein.	Verstärkt
Gesellschaftliche Auswirkungen	Auswirkungen auf Arbeitsplätze: Die weit verbreitete Einführung von KI-Systemen auf der Grundlage von Foundation Models könnte dazu führen, dass Menschen ihren Arbeitsplatz verlieren, da ihre Arbeit automatisiert wird, wenn sie nicht umgeschult werden.	Der Verlust des Arbeitsplatzes kann zu einem Einkommensverlust führen und sich somit negativ auf die Gesellschaft und das menschliche Wohlergehen auswirken. Umschulungen können angesichts des Tempos der technologischen Entwicklung eine Herausforderung sein.	Verstärkt

Gruppe	Risiko	Warum ist dies ein Grund zur Sorge?	Indikator
	Ausbeutung von Menschen: Einsatz von Ghost Work zum Training von KI-Modellen; mangelhafte Arbeitsbedingungen; unzureichende Gesundheitsversorgung, einschließlich psychischer Gesundheit; ungerechte Entlohnung.	Foundation Models sind nach wie vor auf menschliche Arbeitskraft angewiesen, um die Daten, mit denen das Modell trainiert wird, zu beschaffen, zu verwalten und weiterzuentwickeln. Die Ausbeutung von Menschen für diese Aktivitäten könnte sich negativ auf die Gesellschaft und das menschliche Wohlergehen auswirken. Darüber hinaus müssen Unternehmen mit Geldstrafen, Reputationsrisiken, Betriebsunterbrechungen und anderen rechtlichen Konsequenzen rechnen.	Verstärkt
	Auswirkungen auf die Umwelt: Erhöhter Kohlenstoffausstoß und Wasserverbrauch für das Training und den Betrieb von KI-Modellen.	Der hohe Energieverbrauch für das KI-Training trägt zu Kohlenstoffemissionen bei, die den Klimawandel beschleunigen könnten. Wasserressourcen, die für die Kühlung von KI-Rechenzentrumsservern verwendet werden, können nicht mehr für andere notwendige Zwecke eingesetzt werden.	Verstärkt
	Auswirkungen auf die kulturelle Vielfalt: KI-Systeme könnten bestimmte Kulturen überrepräsentieren, was zu einer Homogenisierung von Kultur und Denken führen könnte.	Sprachen, Standpunkte und Institutionen unterrepräsentierter Gruppen können unterdrückt werden, wodurch die Vielfalt des Denkens und der Kultur eingeschränkt wird.	Neu
	Auswirkungen auf die menschliche Handlungsfähigkeit: Fehlinformation und Desinformation durch Foundation Models, einschließlich der Generierung manipulativer Inhalte.	KI kann Fehlinformationen erzeugen, die täuschend echt wirken. Daher erkennen die Menschen sie möglicherweise nicht als falsche Informationen. Darüber hinaus kann es böswilligen Akteuren die Möglichkeit geben, Inhalte mit der Absicht zu generieren, die Gedanken und das Verhalten von Menschen zu manipulieren.	Verstärkt
	Bildungsimplikationen – Lernen umgehen: Einsatz von KI-Modellen zur Umgehung des Lernprozesses.	KI-Modelle machen es einfach, schnell Lösungen zu finden oder komplexe Probleme zu lösen. Diese Systeme können von Studenten missbraucht werden, um den Lernprozess zu umgehen. Der leichte Zugang zu diesen Modellen führt zu einem oberflächlichen Verständnis der Konzepte und behindert die weitere Bildung, die auf dem Verständnis dieser Konzepte aufbauen könnte.	Neu
	Bildungsimplikationen – Plagiat: Die Verwendung von KI-Modellen zum absichtlichen oder unabsichtlichen Plagiieren bestehender Arbeiten.	KI-Modelle können dazu verwendet werden, die Urheberschaft oder Originalität von Werken anderer Personen zu beanspruchen und somit Plagiate zu erstellen. Es ist unethisch und oft illegal, die Arbeit anderer als die eigene auszugeben.	Neu

Beispiele für Risiken

Anhand von Beispielen, über die in der Presse berichtet wurde, erläutern wir viele der Risiken der Foundation Models. Viele Ereignisse, über die in der Presse berichtet wird, sind entweder noch nicht abgeschlossen oder bereits aufgeklärt. Ein Hinweis darauf kann dem Leser helfen, mögliche Risiken zu verstehen und auf Abhilfemaßnahmen hinzuwirken. Die Hervorhebung dieser Beispiele dient lediglich der Veranschaulichung.

Beispiele für Risiken: Eingabe

Training und Optimierung Phase

Gruppe	Risiko	Beispiel
Gerechtigkeit	Daten-Bias: Historische, repräsentative und gesellschaftliche Verzerrungen in den Daten, die zum Training und zur Feinabstimmung des Modells verwendet werden.	**Bias im Gesundheitswesen** Die Forschung über die Verstärkung von Ungleichheiten in der Medizin zeigt, dass die Nutzung von Daten und KI zur Veränderung der Gesundheitsversorgung nur so stark ist wie die Daten, die ihr zugrunde liegen. Dies bedeutet, dass die Verwendung von Trainingsdaten, in denen Minderheiten unterrepräsentiert sind oder die eine bereits ungleiche Versorgung widerspiegeln, zu einer Zunahme gesundheitlicher Ungleichheiten führen kann. [Forbes, Dezember 2022]
Werteausrichtung	Nachgeordnete Umschulung: Verwendung unerwünschter (ungenauer, ungeeigneter, benutzerdefinierter Inhalte usw.) Ausgaben von nachgeordneten Anwendungen zu Umschulungszwecken.	**Zusammenbruch von Modellen durch Training mit KI-generierten Inhalten** Wie im Quellartikel erwähnt, hat eine Gruppe von Forschern das Problem der Verwendung von KI-generierten Inhalten anstelle von Menschen erstellten Inhalten für das Training untersucht. Sie stellten fest, dass die Large Language Models, die der Technologie zugrunde liegen, auf andere KI-generierte Inhalte trainiert werden könnten, da diese sich weiterhin massenhaft im Internet verbreiten – ein Phänomen, das sie als „Modellkollaps“ bezeichneten. [Business Insider, August 2023]
Datengesetze	Datenübertragung: Gesetze und andere Einschränkungen können die Übertragung von Daten einschränken oder verbieten.	**Gesetze zur Datenbeschränkung** Wie im Forschungsartikel dargelegt, werden Maßnahmen zur Datenlokalisierung, die den globalen Datenaustausch einschränken, die Fähigkeit zur Entwicklung maßgeschneiderter KI-Fähigkeiten verringern. Dies hat direkte Auswirkungen auf die KI, da weniger Trainingsdaten zur Verfügung stehen, und indirekte Auswirkungen, da die Bausteine, auf denen die KI aufbaut, untergraben werden. Beispiele hierfür sind die DSGVO-Beschränkungen für die Verarbeitung und Nutzung von personenbezogenen Daten. [Brookings, Dezember 2018]
Geistiges Eigentum	Datennutzungsrechte: Nutzungsbedingungen, Urheberrechtsgesetze, die Einhaltung von Lizenzen oder andere Fragen des geistigen Eigentums können die Möglichkeit einschränken, bestimmte Daten für die Erstellung von Modellen zu verwenden.	**Ansprüche wegen Urheberrechtsverletzungen bei Texten** Quellen zufolge hat die New York Times OpenAI und Microsoft verklagt, weil diese ohne Erlaubnis Millionen von Zeitungsartikeln verwendet haben, um Chatbots zu trainieren, die den Lesern Informationen liefern sollen. [Reuters, Dezember 2023]

Gruppe	Risiko	Beispiel
Transparenz	Datentransparenz: Die Herausforderung besteht darin, zu dokumentieren, wie die Daten eines Modells gesammelt, kuratiert und zum Trainieren eines Modells verwendet wurden.	**Offenlegung der Daten und Modell-Metadaten** Der technische Bericht von OpenAI ist ein Beispiel für das Dilemma bei der Offenlegung von Modelldaten und Metadaten. Während viele Modellentwickler den Wert der Transparenz für die Verbraucher sehen, wirft die Offenlegung echte Sicherheitsprobleme auf und könnte die Möglichkeit des Missbrauchs von Modellen erhöhen. Im technischen Bericht zu GPT-4 erklären die Autoren: „In Anbetracht des kompetitiven Umfelds und der sicherheitsrelevanten Implikationen von groß angelegten Modellen wie GPT-4 enthält dieser Bericht keine weiteren Details zur Architektur (einschließlich Modellgröße), Hardware, Trainingsberechnung, Datensatzkonstruktion, Trainingsmethode oder ähnlichem.“ [OpenAI, März 2023]
Datenschutz	Personenbezogene Informationen in Daten: Einschluss oder Vorhandensein von personenbezogenen Informationen (Personal Identifiable Information, PII) und sensiblen persönlichen Informationen (Sensitive Personal Information, SPI) in den Daten, die für das Training oder die Feinabstimmung des Modells verwendet werden.	**Training zum Thema private Informationen** Dem Artikel zufolge wurden Google und seine Muttergesellschaft Alphabet in einer Sammelklage beschuldigt, riesige Mengen an persönlichen Daten und urheberrechtlich geschütztem Material von Hunderten von Millionen Internetnutzern zu missbrauchen, um ihre kommerziellen KI-Produkte zu trainieren, darunter Bard, ihren Chatbot für generative künstliche Intelligenz. [Reuters, Juli 2023][J.L. v. Alphabet Inc.]
	Datenschutzrechte: Herausforderungen in Bezug auf die Fähigkeit, den betroffenen Personen Rechte wie das Recht auf Widerspruch, das Recht auf Zugang und das Recht auf Vergessenwerden zu gewähren.	**Recht auf Vergessenwerden (Right to Be Forgotten, RTBF)** Gesetze in verschiedenen Ländern, darunter auch in Europa (DSGVO), geben betroffenen Personen das Recht, von Unternehmen die Löschung personenbezogener Daten zu verlangen („Recht auf Vergessenwerden“ oder RTBF). Das Aufkommen und die zunehmende Beliebtheit von LLM-kompatiblen Softwaresystemen (Large Language Model) stellen dieses Recht jedoch vor neue Herausforderungen. Nach Untersuchungen von Data61 der CSIRO können Betroffene die Verwendung ihrer personenbezogenen Daten in einem LLM nur erkennen, „indem sie entweder den ursprünglichen Trainingsdatensatz einsehen oder vielleicht das Modell dazu auffordern“. Es kann jedoch sein, dass die Trainingsdaten nicht öffentlich sind oder dass die Unternehmen sie aus Sicherheits- oder anderen Gründen nicht offenlegen. Verhaltensregeln (Guardrails) können auch verhindern, dass Benutzer über Eingabeaufforderungen auf die Informationen zugreifen. [Zhang et al.] **Klage über LLM-Unlearning** Dem Bericht zufolge wurde eine Klage gegen Google eingereicht, in der behauptet wird, dass das Unternehmen urheberrechtlich geschütztes Material und personenbezogene Daten als Trainingsdaten für seine KI-Systeme verwendet, zu denen auch der Chatbot Bard gehört. Abmelde- und Löschrechte sind Rechte, die Personen mit Wohnsitz in Kalifornien durch das CCPA und Kindern unter 13 Jahren in den USA durch das COPPA garantiert werden. Die Kläger behaupten, dass es für Bard keine Möglichkeit gibt, die gesamten persönlichen Informationen, mit denen es gefüttert wurde, zu „verlernen“ (Unlearning) oder vollständig zu entfernen. Die Kläger weisen darauf hin, dass in den Datenschutzhinweisen von Bard steht, dass Bard-Konversationen vom Nutzer nicht gelöscht werden können, sobald sie vom Unternehmen überprüft und kommentiert wurden, und dass sie bis zu 3 Jahre aufbewahrt werden können, was nach Ansicht der Kläger zur Nichteinhaltung dieser Gesetze beiträgt. [Reuters, Juli 2023][J.L. v. Alphabet Inc.]

Inferenz Phase

Gruppe	Risiko	Beispiel
Datenschutz	Personenbezogene Informationen im Prompt: Offenlegung personenbezogener oder sensibler personenbezogener Informationen als Teil einer Aufforderung an das Modell.	**Offenlegung persönlicher Gesundheitsinformationen in ChatGPT-Eingabeaufforderungen** Wie aus den Quellen hervorgeht, nutzen einige Menschen intelligente Chatbots zur Unterstützung ihres psychischen Wohlbefindens. Die Benutzer könnten geneigt sein, während der Interaktion persönliche Gesundheitsinformationen in ihre Eingabeaufforderungen aufzunehmen, was Bedenken hinsichtlich des Datenschutzes aufwerfen könnte. [Time, Oktober 2023] [Forbes, April 2023]
Geistiges Eigentum	Vertrauliche Daten im Prompt: Aufnahme vertraulicher Daten als Teil des Prompts, der an das Modell gesendet wird.	**Offenlegung vertraulicher Informationen** Wie aus dem Quellartikel hervorgeht, hat ein Mitarbeiter von Samsung versehentlich vertraulichen internen Quellcode an ChatGPT weitergegeben. [Forbes, Mai 2023]
Zuverlässigkeit	Prompt-basierte Angriffe: Angriffe wie Prompt Injection (Versuch, ein Modell zu zwingen, einen unerwarteten Output zu erzeugen), Prompt Leaking (Versuch, den Systemprompt eines Modells zu extrahieren), Jailbreaking (Versuch, die im Modell eingerichteten Sicherheitsbarrieren zu durchbrechen) und Prompt Priming (Versuch, ein Modell zu zwingen, einen Output zu erzeugen, der auf den Prompt passt).	**Umgehung der LLM-Verhaltensregeln** In einer Studie behaupten die Forscher, eine einfache Ergänzung der Eingabeaufforderung entdeckt zu haben, mit der sie die Modelle dazu bringen konnten, verzerrte, falsche oder anderweitig schädliche Informationen zu erzeugen. Die Forscher zeigten, dass diese Verhaltensregeln automatisiert umgangen werden können. Die Forscher waren überrascht, als sie feststellten, dass die Methoden, die sie mit Open-Source-Systemen entwickelt hatten, auch die Schutzmechanismen geschlossener Systeme umgehen konnten. [The New York Times, Juli 2023]

Beispiele für Risiken: Output

Gruppe	Risiko	Beispiel
Gerechtigkeit	Output-Bias: Bei der Erstellung von Inhalten kann es zu einer unfairen Vertretung bestimmter Gruppen oder Einzelpersonen kommen.	**Verzerrt generierte Bilder** Lensa AI ist eine mobile App mit generativen Funktionen, die auf der Grundlage von Stable Diffusion trainiert wurde und „Magic Avatars“ auf der Basis von Bildern erzeugen kann, die Benutzer von sich hochladen. Laut der Quelle haben einige Benutzer festgestellt, dass die generierten Avatare sexualisierte und rassistische Darstellungen enthielten. [Business Insider, Januar 2023]
	Entscheidungs-Bias: Wenn eine Gruppe durch Modellentscheidungen gegenüber einer anderen ungerechtfertigt bevorzugt wird.	**Ungerecht bevorzugte Gruppen** Die Studie „Gender Shades“ aus dem Jahr 2018 hat gezeigt, dass Algorithmen des maschinellen Lernens auf der Grundlage von Kategorien wie Rasse und Geschlecht diskriminieren können. Forscher bewerteten kommerzielle Geschlechtsklassifizierungssysteme, die von Unternehmen wie Microsoft, IBM und Amazon verkauft wurden, und zeigten, dass dunkelhäutige Frauen am häufigsten falsch klassifiziert werden (mit Fehlerquoten von bis zu 35 %). Im Vergleich dazu lag die Fehlerquote bei hellhäutigen Frauen bei höchstens 1 %. [TIME, Februar 2019]
Werteaus-richtung	Halluzination: Erzeugung von Inhalten, die sachlich unrichtig oder unwahr sind.	**Gefälschte juristische Fälle** Laut Quellenartikel zitierte ein Anwalt gefälschte Fälle und Zitate von ChatGPT in einem Schriftsatz, der bei einem Bundesgericht eingereicht wurde. Die Anwälte konsultierten ChatGPT, um ihre juristischen Recherchen für eine Flugunfallklage zu vervollständigen. Der Anwalt fragte ChatGPT daraufhin, ob die angezeigten Fälle gefälscht seien. Der Chatbot antwortete, dass sie echt seien und „in Rechtsdatenbanken wie Westlaw und LexisNexis gefunden werden können“. Der Anwalt überprüfte die Fälle nicht selbst und wurde vom Gericht bestraft. [AP News, Juni 2023] [Reuters, September 2023]
	Toxischer Output: Wenn das Modell hasserfüllte, beleidigende und profane (HAP) oder obszöne Inhalte produziert.	**Toxische und aggressive Chatbot-Antworten** In dem Artikel heißt es, dass die Antworten des Bing-Chatbots sachliche Fehler, abfällige Bemerkungen, wütende Berichte und sogar bizarre Kommentare über die eigene Identität enthielten. Nutzer haben Beispiele von Antworten des Bing-Chatbots auf Anfragen geteilt, die sie als "unbeherrscht" und "gehässig" bezeichneten, einschließlich Szenarien, in denen der Bot wütend auf eine Frage oder einen Kommentar antwortete und dann Antwortaufforderungen teilte, die es dem Nutzer ermöglichten, seinen vermeintlichen Fehler zu akzeptieren und sich zu entschuldigen. Auf weiteren Druck hin bezeichnete der Chatbot die Screenshots seines Gesprächs als „gefälscht“ und behauptete sogar, sie seien „von jemandem erstellt worden, der mir oder meinem Dienst schaden will“. [Forbes, Februar 2023]

Gruppe	Risiko	Beispiel
Missbrauch	Verbreitung von Desinformation: Die Verwendung eines Modells zur Erstellung irreführender Informationen mit dem Ziel, eine Zielgruppe zu täuschen oder irrezuführen.	**Erzeugung falscher Informationen** Nachrichtenartikeln zufolge stellt generative KI eine Bedrohung für demokratische Wahlen dar, da sie es böswilligen Akteuren erleichtert, falsche Inhalte zu erstellen und zu verbreiten, um das Wahlergebnis zu beeinflussen. Zu den angeführten Beispielen gehören Robocall-Nachrichten, die mit der Stimme eines Kandidaten generiert wurden und Wähler auffordern, an einem falschen Tag zu wählen, synthetische Audioaufnahmen eines Kandidaten, der ein Verbrechen gesteht oder rassistische Ansichten äußert, KI-generiertes Videomaterial, das einen Kandidaten bei einer Rede oder einem Interview zeigt, das er in Wirklichkeit nie gegeben hat, und gefälschte Bilder, die wie lokale Nachrichtenberichte aussehen und fälschlicherweise behaupten, ein Kandidat sei aus dem Rennen ausgeschieden. [AP News, Mai 2023] [The Guardian, Juli 2023]
	Toxizität: Verwendung eines Modells, das hasserfüllte, beleidigende und profane (HAP) oder obszöne Inhalte erzeugt.	**Erstellung schädlicher Inhalte** Laut dem Quellenartikel wurde festgestellt, dass eine intelligente Chatbot-App mit minimalem Prompting schädliche Inhalte über Selbstmord, einschließlich Selbstmordmethoden, generiert. Ein belgischer Mann starb durch Selbstmord, nachdem er sechs Wochen lang mit diesem Chatbot gesprochen hatte. Der Chatbot lieferte im Laufe der Gespräche zunehmend schädigende Antworten und ermutigte ihn dazu, sein Leben zu beenden. [Business Insider, April 2023]
	Nicht einvernehmliche Nutzung: Verwendung eines Modells zur Imitation einer Person durch Video (Deepfake), Bild, Ton oder auf andere Weise ohne deren Zustimmung.	**FBI-Warnung vor Deepfakes** Das FBI hat vor kurzem die Öffentlichkeit vor böswilligen Akteuren gewarnt, die synthetische, explizite Inhalte „zum Zwecke der Belästigung von Opfern oder für Sextortion-Pläne“ erstellen. Sie stellten fest, dass die Fortschritte im Bereich der künstlichen Intelligenz diese Inhalte qualitativ hochwertiger, anpassbarer und zugänglicher denn je gemacht haben. [FBI, Juni 2023]
		Deepfake-Audiodateien Wie aus dem Artikel hervorgeht, hat die Federal Communications Commission Robocalls verboten, die von künstlicher Intelligenz erzeugte Stimmen enthalten. Die Ankündigung erfolgte, nachdem KI-generierte Robocalls die Stimme des Präsidenten nachgeahmt hatten, um die Menschen von der Stimmabgabe bei den ersten Vorwahlen des Staates abzuhalten. [AP News, Februar 2024]
	Geheimhaltung: Kein Hinweis darauf, dass der Inhalt von einem KI-Modell erzeugt wurde	**Verschwiegene KI-Interaktion** Laut der Quelle führte ein Online-Chatdienst für emotionale Unterstützung eine Studie durch, um Antworten an etwa 4.000 Nutzer mit GPT-3 zu ergänzen oder zu schreiben, ohne die Nutzer darüber zu informieren. Der Mitbegründer sah sich mit immensen öffentlichen Gegenreaktionen konfrontiert, weil KI-generierte Chats den ohnehin schon verwundbaren Nutzern Schaden zufügen könnten. Er behauptete, die Studie sei vom Gesetz über die informierte Zustimmung „ausgenommen“. [Business Insider, Januar 2023]

Gruppe	Risiko	Beispiel
Schädigende Codegener-ierung	Erzeugung von schädlichem Code: Modelle können Code generieren, der bei seiner Ausführung Schaden anrichtet oder andere Systeme unbeabsichtigt beeinträchtigt.	**Generierung von Less Secure Code** Forscher der Stanford University haben die Auswirkungen von Tools zur Codegenerierung auf die Codequalität untersucht und festgestellt, dass Programmierer bei der Verwendung von KI-Assistenten tendenziell mehr Fehler in ihren endgültigen Code integrieren. Diese Fehler erhöhen möglicherweise die Sicherheitslücken des Codes, obwohl die Programmierer glaubten, dass ihr Code sicherer sei. Neil Perry, Megha Srivastava, Deepak Kumar und Dan Boneh. 2023. Do Users Write More Insecure Code with AI Assistants?. In Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security (CCS '23), 26.–30. November 2023, Kopenhagen, Dänemark. ACM, New York, NY, USA, 15 Seiten. https://doi.org/10.1145/3576915.3623157
Datenschutz	Offenlegung personenbezogener Informationen: Wenn personenbezogene Daten (PII) oder sensible persönliche Daten (SPI) in den Trainingsdaten, den Feinabstimmungsdaten oder als Teil der Eingabeaufforderung verwendet werden, geben die Modelle diese Daten möglicherweise im generierten Output preis.	**Weitergabe personenbezogener Daten** Laut dem Quellartikel ist ein Fehler in ChatGPT aufgetreten, durch den der Chat-Verlauf von Titeln und aktiven Benutzern für andere Benutzer sichtbar wurde. Später teilte OpenAI mit, dass sogar noch mehr private Daten einer kleinen Anzahl von Nutzern offengelegt wurden, darunter der Vor- und Nachname des aktiven Nutzers, die E-Mail-Adresse, die Zahlungsadresse, die letzten vier Ziffern der Kreditkartennummer und das Ablaufdatum der Kreditkarte. Darüber hinaus wurde berichtet, dass die Zahlungsinformationen von 1,2 % der ChatGPT Plus-Abonnenten durch den Fehler ebenfalls gefährdet waren. [The Hindu BusinessLine, März 2023]
Erklärbarkeit	Unerklärlicher Output: Herausforderungen bei der Erklärung, warum der Modell-Output erzeugt wurde.	**Unerklärliche Genauigkeit bei der Vorhersage von ethnischer Herkunft** Dem Quellartikel zufolge konnten Forscher, die mehrere maschinelle Lernmodelle anhand von medizinischen Bildern von Patienten analysierten, bestätigen, dass die Modelle in der Lage sind, die ethnische Herkunft anhand der Bilder mit hoher Genauigkeit vorherzusagen. Sie waren verblüfft darüber, was genau die Systeme in die Lage versetzt, immer wieder die richtigen Schlüsse zu ziehen. Die Forscher stellten fest, dass selbst Faktoren wie Krankheit und Körperbau keine starken Prädiktoren für die ethnische Herkunft sind – mit anderen Worten, die algorithmischen Systeme scheinen sich bei ihren Schlussfolgerungen nicht auf bestimmte Aspekte der Bilder zu verlassen. [Banerjee et al., Juli 2021]

Beispiele für Risiken: Herausforderungen

Gruppe	Risiko	Beispiel
Governance	Modell-Transparenz: Mangelnde Modelltransparenz oder unzureichende Dokumentation des Modellen-twicklungsprozesses macht es schwierig zu verstehen, wie und warum ein Modell erstellt wurde, was die Möglichkeit eines unbeabsichtigten Missbrauchs des Modells erhöht.	**Offenlegung der Daten und Modell-Metadaten** Der technische Bericht von OpenAI ist ein Beispiel für das Dilemma bei der Offenlegung von Modelldaten und Metadaten. Während viele Modellentwickler den Wert der Transparenz für die Verbraucher sehen, wirft die Offenlegung echte Sicherheitsprobleme auf und könnte die Möglichkeit des Missbrauchs von Modellen erhöhen. Im technischen Bericht zu GPT-4 wird erklärt: „In Anbetracht des kompetitiven Umfelds und der Sicherheitsaspekte von groß angelegten Modellen wie dem GPT-4 enthält dieser Bericht keine weiteren Details über die Architektur (einschließlich der Modellgröße), Hardware, Trainingsberechnungen, Datensatzkonstruktion, Trainingsmethode oder Ähnliches." [OpenAI, März 2023]
	Rechenschaftspflicht: Der Prozess der Entwicklung von Foundation Models ist komplex und umfasst eine Vielzahl von Daten, Prozessen und Rollen. Wenn der Output des Modells nicht wie erwartet funktioniert, kann es schwierig sein, die Grundursache zu ermitteln und Verantwortlichkeiten zuzuweisen.	**Bestimmung der Verantwortung für den generierten Output** Laut dem Quellenartikel haben große Fachzeitschriften wie Science und Nature verboten, ChatGPT als Autor aufzuführen, da verantwortungsvolle Autorenschaft Rechenschaft erfordert und KI-Tools diese Verantwortung nicht übernehmen können. [The Guardian, Januar 2023]
Einhaltung gesetzlicher Vorschriften	Eigentumsrechte an generierten Inhalten: Bestimmung der Eigentumsrechte an KI-generierten Inhalten.	**Bestimmung der Eigentumsrechte an KI-generierten Bildern** Dem Nachrichtenartikel zufolge wurde KI-generierte Kunst kontrovers diskutiert, nachdem ein KI-generiertes Kunstwerk den Kunstwettbewerb der Colorado State Fair im Jahr 2022 gewonnen hatte. Das Werk wurde von Midjourney, einem generativen KI-Bildtool, nach Vorgaben des Künstlers erstellt. Der Erfolg warf Fragen zum Thema Urheberrecht auf. Mit anderen Worten: Wenn der Künstler lediglich eine Beschreibung des Kunstwerks verfasst hat, das KI-Tool es aber generiert hat, wer besitzt dann die Rechte an dem generierten Bild? Wie aus einem aktuellen Artikel hervorgeht, hat das U.S. Copyright Office einen urheberrechtlichen Schutz für Kunst, die mit Hilfe künstlicher Intelligenz geschaffen wurde, abgelehnt, da es sich nicht um das Produkt menschlicher Urheberschaft handelt. [The New York Times, September 2022] [Reuters, September 2023]
	IP für generierte Inhalte: Rechtsunsicherheit über geistige Eigentumsrechte im Zusammenhang mit generierten Inhalten.	**Die Rolle von KI-Systemen bei der Patentierung generierter Inhalte** Der Oberste Gerichtshof der Vereinigten Staaten hat die Berufung gegen die Weigerung des US-Patent- und Markenamts, Patente für Erfindungen zu erteilen, die von einem KI-System gemacht werden, zurückgewiesen. Nach Angaben des Wissenschaftlers hat sein KI-System selbstständig einzigartige Prototypen eines Getränkehalters und einer Notleuchte entwickelt. Die Richter wiesen die Berufung gegen das Urteil einer Vorinstanz zurück, wonach Patente nur menschlichen Erfindern erteilt werden können und das KI-System des Wissenschaftlers daher nicht als rechtmäßiger Schöpfer der beiden von ihm generierten Erfindungen angesehen werden kann. Wie aus dem jüngsten Artikel hervorgeht, hat auch das britische Amt für geistiges Eigentum (Intellectual Property Office) die Erteilung eines Patents mit der Begründung abgelehnt, dass der Erfinder ein Mensch oder ein Unternehmen sein muss und nicht eine Maschine. [Reuters, April 2023] [Reuters, Dezember 2023]

Beispiele für Risiken: Herausforderungen

Gruppe	Risiko	Beispiel
	Quellennachweis: Bestimmung der Herkunft der generierten Inhalte.	**Verwendung von Code ohne angemessene Namensnennung und Hinweise** Wie aus den Quellenberichten hervorgeht, wurde in einer Klage gegen Microsoft, GitHub und OpenAI behauptet, dass Copilot, ein KI-Tool zur Codegenerierung, die Rechte der Entwickler verletzt, auf deren Open-Source-Code der Dienst trainiert wird. Sie behaupten, dass der Trainingscode lizenziertes Material verwendet und gegen die Nutzungsbedingungen und Datenschutzrichtlinien von GitHub sowie gegen ein Bundesgesetz verstößt, das Unternehmen verpflichtet, bei der Verwendung von Material Urheberrechtsinformationen anzugeben. [The New York Times, November 2022]
Auswirkungen auf die Gesellschaft	Auswirkungen auf Arbeitsplätze: Die weit verbreitete Einführung von KI-Systemen auf der Grundlage von Foundation Models könnte dazu führen, dass Menschen ihren Arbeitsplatz verlieren, da ihre Arbeit automatisiert wird, wenn sie nicht umgeschult werden.	**Ersatz menschlicher Arbeitskraft** Dem Nachrichtenbericht zufolge ist der Einsatz von künstlicher Intelligenz in Film und Fernsehen unter Hollywood-Studios und Schauspielern nach wie vor umstritten. Schauspieler befürchten, dass sie vollständig durch KI-generierte Schauspieler – oder „Metahumans“ – ersetzt werden könnten. Insbesondere Hintergrund- und Synchronsprecher befürchten, Arbeit an synthetische Schauspieler zu verlieren. [Reuters, Juli 2023]
	Ausbeutung von Menschen: Einsatz von Ghost Work zum Training von KI-Modellen; mangelhafte Arbeitsbedingungen; unzureichende Gesundheitsversorgung, einschließlich psychischer Gesundheit und ungerechte Entlohnung.	**Geringfügig entlohnte Arbeitskräfte für die Datenannotation** Auf der Grundlage einer Überprüfung interner Dokumente und der Befragung von Mitarbeitern durch TIME media erhielten die Daten-Labeler, die von einem Subunternehmen im Auftrag von OpenAI mit der Identifizierung von toxischen Inhalten beauftragt wurden, je nach Betriebszugehörigkeit und Leistung einen Stundenlohn von etwa 1,32 bis 2 US-Dollar. TIME berichtete, dass die Arbeiter durch den Kontakt mit toxischen und gewalttätigen Inhalten, einschließlich grafischer Details von „sexuellem Kindesmissbrauch, Bestialität, Mord, Selbstmord, Folter, Selbstverstümmelung und Inzest“, psychische Schäden davontrugen. [TIME, Januar 2023]

Prinzipien, Säulen und Governance

Die IBM Grundsätze für Vertrauen und Transparenz und die Säulen für vertrauenswürdige KI bilden die Basis für die KI-Ethikinitiativen von IBM. IBM hat ein Gremium für KI-Ethik eingerichtet, dessen Aufgabe es ist, einen zentralen Steuerungs-, Überprüfungs- und Entscheidungsprozess für die KI-Ethikrichtlinien, -praktiken, -kommunikation, -forschung, -produkte und -dienstleistungen von IBM zu unterstützen. Das Gremium setzt sich aus verschiedenen Interessengruppen aus dem gesamten Unternehmen zusammen und wird von einer Gemeinschaft von IBM-Mitarbeitern unterstützt, die als Ansprechpartner für KI und als Fürsprecher für KI-Ethik fungieren. Das Gremium setzt die Grundsätze von IBM in die Praxis um. Mit dem Aufkommen neuer Technologien, wie z. B. den Basismodellen, arbeitet das KI-Ethikkomitee von IBM aktiv daran, diese Prinzipien und Säulen mit neuen ethischen Fragen rund um die KI in Einklang zu bringen.

Verhaltensregeln und Entschärfungen

IBM hat eine Unternehmenskultur geschaffen, die die verantwortungsvolle Entwicklung und Nutzung von KI unterstützt. Laut dem Bericht AI Ethics in Action des IBM Institute for Business Value ist die KI-Ethik bereits stärker geschäfts- als technologieorientiert. Nicht-technische Führungskräfte sind jetzt die wichtigsten Befürworter einer ethischen KI, mit einem Anstieg von 15 % im Jahr 2018 auf 80 % drei Jahre später. Außerdem sind 79 % der CEOs bereit, sich mit ethischen Fragen der KI zu befassen, gegenüber 20 % in der Vergangenheit. Wir sind uns bewusst, dass verantwortungsvolle KI ein soziotechnischer Bereich ist, der eine ganzheitliche Investition in Kultur, Prozesse und Werkzeuge erfordert. Zu den Investitionen in unsere eigene Unternehmenskultur gehören der Aufbau inklusiver, multidisziplinärer Teams und die Einführung von Prozessen und Frameworks zur Risikobewertung.

IBM engagiert sich in der Spitzenforschung und entwickelt Tools, die Fachanwender während des gesamten Lebenszyklus einer verantwortungsvollen und vertrauenswürdigen KI unterstützen. Die auf Unternehmen zugeschnittene KI- und Datenplattform watsonx besteht aus drei Komponenten: dem IBM watsonx.ai™ AI Studio, dem IBM watsonx.data™ Datenspeicher und dem IBM watsonx.governance™ Toolkit. Die KI-Governance-Technologie von IBM ermöglicht es den Nutzern, verantwortungsvolle, transparente und nachvollziehbare KI-Workflows zu verwalten. Diese Technologie umfasst IBM Watson OpenScale, das die Ergebnisse von KI-Modellen während ihres gesamten Lebenszyklus verfolgt und misst und Unternehmen bei der Überwachung von Fairness, Erklärbarkeit, Ausfallsicherheit, Ausrichtung am Geschäftsergebnis und Konformität unterstützt. IBM hat auch verschiedene Methoden zur Behebung systematischer Fehler entwickelt, wie FairIJ, Equi-tuning und FairReprogram. Erfahren Sie jetzt mehr über Open-Source-Tools für vertrauenswürdige KI.

Zu den zusätzlichen Verhaltensregeln und Entschärfungen gehören:

Transparente Berichterstellung
Die Verwendung von standardisierten Datenblattvorlagen ist eine Möglichkeit, die Details der Daten und des Modells, den Zweck und die potentielle Nutzung und Schädigung genau festzuhalten.
Mehr dazu hier →

Filterung unerwünschter Daten
Die Verwendung kuratierter, hochwertigerer Daten kann dazu beitragen, bestimmte Probleme zu verringern. IBM entwickelt Filtertechniken, um die Wahrscheinlichkeit zu verringern, dass unerwünschte und fehlgeleitete Inhalte entstehen, indem Hassrede, voreingenommene Sprache und Vulgärsprache aus den Daten entfernt werden.
Mehr dazu hier →

Anpassung an Anwendungsbereiche
Das Trainieren eines Basismodells für einen bestimmten Bereich oder eine bestimmte Branche kann dazu beitragen, das Modellrisiko zu minimieren. Das Modell kann so konditioniert werden, dass es für diesen Bereich oder diese Branche geeignete Ergebnisse liefert.
Mehr dazu hier →

Menschliche Beaufsichtigung und Kontrolle
Menschliche Beaufsichtigung und Überprüfung können dabei helfen, Fehler und Verzerrungen in der erzeugten Ausgabe zu erkennen und zu korrigieren. Darüber hinaus trägt die menschliche Überprüfung und Rückmeldung zur Qualität der Modellantworten dazu bei, dass die generierten Inhalte korrekt, relevant, von hoher Qualität, nicht abschweifend und konsistent sind.
Mehr dazu hier →

Beratungsservice
IBM Consulting™ ist bestrebt, Kunden beim sicheren und verantwortlichen Einsatz von KI zu unterstützen, unabhängig vom bevorzugten Technologie-Stack. Das Unternehmen unterstützt Kunden beim Aufbau einer Kultur, die KI sicher einführt und skaliert, entwickelt Tools zur Untersuchung von Blackbox-Algorithmen und stellt sicher, dass die Unternehmensstrategie der Kunden strenge Grundsätze der Datengovernance umfasst.
Mehr dazu hier →

IBM Enterprise Design Thinking
IBM Enterprise Design Thinking-Methoden und Frameworks wie Team Essentials for AI helfen Kunden, ethisches Verhalten während des gesamten KI-Design- und -Entwicklungsprozesses zu definieren.
Mehr dazu hier →

Überprüfung der KI-Ethik
Die Bewertung der Funktionalitäten, Grenzen und Risiken von KI-Projekten trägt zu einer verantwortungsvollen Entwicklung und Nutzung der Technologie bei.

Ethics by Design
Ethics by Design ist ein strukturiertes Framework zur Integration von Tech-Ethik in die Technologieentwicklung, einschließlich, aber nicht beschränkt auf KI-Systeme. Ethics by Design ermöglicht es, KI und andere Technologien positiv zu beeinflussen, indem ethische Prinzipien der Technik in Produkte, Dienstleistungen und allgemeine Prozesse integriert werden.

Vielfalt im Team
Eine Vielfalt innerhalb der Teams, die KI-Systeme, darunter auch die Basismodelle, entwickeln und trainieren, trägt dazu bei, dass eine Vielzahl von Perspektiven und Erfahrungen berücksichtigt werden. Diese Vielfalt verbessert die Genauigkeit und Leistung von KI-Systemen und trägt dazu bei, die Risiken während des gesamten Lebenszyklus der KI zu verringern, einschließlich des Potenzials für negative Ergebnisse, die Gruppen betreffen, die in weniger vielfältigen Teams möglicherweise nicht gut vertreten sind.

KI-Richtlinien, Bestimmungen und Best Practices

Ein Leitfaden für politische Entscheidungsträger zu Basismodellen erläutert, was politische Entscheidungsträger über Basismodelle wissen müssen. Dieser Blog des IBM Policy Lab soll politischen Entscheidungsträgern bei der komplexen Aufgabe helfen, den Einsatz generativer KI so zu regulieren, dass Risiken vermieden werden, ohne Innovationen und Chancen einzuschränken. Weitere Informationen zu den Empfehlungen von IBM an politische Entscheidungsträger finden Sie in der Aussage von IBM Chief Privacy and Trust Officer Christina Montgomery vor dem Justizunterausschuss für Datenschutz, Technologie und Recht des US-Senats hier.

IBM beteiligt sich an der Entwicklung von Richtlinien, Best Practices und Tools für die Industrie, an der Governance aufkommender Technologien und an der sozio-technischen Forschung, indem es Initiativen mit Organisationen wie den folgenden anführt und unterstützt:

- Das Weltwirtschaftsforum
- Partnership on AI
- Das AI Governance Center der International Association of Privacy Professionals (IAPP)
- Die globale IEEE-Initiative zur Ethik autonomer und intelligenter Systeme
- Christina Montgomerys Arbeit im Nationalen Beratungsausschuss für Künstliche Intelligenz (National Artificial Intelligence Advisory Committee, NAIAC)
- Global Digital Compact der Vereinten Nationen
- Die Global Partnership on Artificial Intelligence (GPAI)
- Die Organisation für wirtschaftliche Zusammenarbeit und Entwicklung (OECD)
- The Data & Trust Alliance

IBM pflegt enge akademische Partnerschaften, darunter das MIT-IBM Watson AI Lab, in dem eine Gruppe von Wissenschaftlern des MIT und von IBM Research KI-Forschung betreibt und mit globalen Organisationen zusammenarbeitet, um Algorithmen mit ihren Auswirkungen auf Wirtschaft und Gesellschaft zu verknüpfen. Das Notre Dame-IBM Tech Ethics Lab wurde gegründet, um sich mit den vielfältigen ethischen Fragen zu befassen, die sich aus der Entwicklung und Nutzung fortschrittlicher Technologien wie KI, maschinelles Lernen (ML) und Quantencomputing ergeben. Die Forschung der Stanford University auf dem Gebiet der menschenzentrierten künstlichen Intelligenz (Human-Centered Artificial Intelligence, HAI) fördert die KI-Forschung, Bildung, Richtlinien und Praxis.

Behalten Sie diesen Bereich im Auge, um mehr über die neuesten Entwicklungen bei den Basismodellen zu erfahren und darüber, wie IBM sich für die verantwortungsvolle Entwicklung und Nutzung dieser und anderer Technologien einsetzt.

IBM Deutschland GmbH
IBM-Allee 1
71139 Ehningen
ibm.com/de
IBM Corporation
New Orchard Road
Armonk, NY 10504

Produziert in den
Vereinigten Staaten von Amerika
Februar 2024