Auf der Grundlage des Hiroshima-Prozesses Internationale:

Leitprinzipien für Organisationen, die fortschrittliche KI-Systeme entwickeln.

Die Entwicklung fortschrittlicher KI-Systeme zielt darauf ab, eine sichere und vertrauenswürdige KI weltweit zu fördern und wird freiwillige Leitlinien für Maßnahmen von Organisationen bereitstellen, die die fortschrittlichste KI entwickeln, einschließlich der fortschrittlichsten Basismodelle und generativen KI-Systeme (im Folgenden "fortschrittliche KI-Systeme").

Unternehmen sollten diese Maßnahmen im Einklang mit einem risikobasierten Ansatz befolgen.

Zu den Organisationen, die diesen Verhaltenskodex unterstützen können, gehören unter anderem Unternehmen aus Wissenschaft, Zivilgesellschaft, Privatwirtschaft und/oder dem öffentlichen Sektor.

Diese nicht erschöpfende Liste von Maßnahmen wird als Reaktion auf die jüngsten Entwicklungen in Sachen fortschrittlicher KI diskutiert und als lebendiges Dokument ausgearbeitet, das die bestehenden KI-Grundsätzen der OECD ergänzen soll.

Es soll dazu beitragen, die Vorteile dieser Technologie zu nutzen und die damit verbundenen Risiken und Herausforderungen zu bewältigen. Unternehmen sollten diese Maßnahmen auf alle Phasen des Lebenszyklus einer KI anwenden. Dies umfasst, falls möglich, die Konzeption, Entwicklung, Einführung und Nutzung fortschrittlicher KI-Systeme.

Dieses Dokument wird bei Bedarf überprüft und aktualisiert, um sicherzustellen, dass es weiterhin seinen Zweck erfüllt und auch die neuesten Entwicklungen dieser sich schnell entwickelnde Technologie berücksichtigt.

Je nach unterschiedlicher Rechtsordnungen können verschiedene Ansätze ergriffen werden, um diese Maßnahmen umzusetzen.

Wir fordern die angesprochenen Organisationen auf, in Absprache mit anderen relevanten Interessengruppen, diese Maßnahmen im Einklang mit einem risikobasierten Ansatz umzusetzen. Neue Regelungen und Regulierungsansätze der Regierung sind hierbei ggf. zu beachten.

Wir verpflichten uns auch, in Absprache mit den OECD, GPAI und andere Interessenträger, zur Einführung von Überwachungsinstrumenten und -mechanismen. Die Organisationen bleiben für die Umsetzung dieser Maßnahmen verantwortlich. Wir ermutigen Organisationen die Entwicklung wirksamer Überwachungsmechanismen zu unterstützen, und durch praktische Erkenntnisse zu verbessern.

Darüber hinaus ermutigen wir Unternehmen interne KI-Regulierungsstrukturen und -Richtlinien einzurichten, einschließlich Selbstbewertungsmechanismen, um einen verantwortungsvollen und rechenschaftspflichtigen Ansatz für die Umsetzung dieser Maßnahmen in der KI-Entwicklung zu schaffen.

Während der Konzeption, Entwicklung und dem Einsatz von KI-Systemen sollten Organisationen zu jeder Zeit die Grundsätze der Rechtsstaatlichkeit, Menschenrechte, eines ordnungsgemäßen Verfahrens, Vielfalt, Fairness und Nichtdiskriminierung,

Demokratie und Menschenorientierung achten, um die Chancen dieser Innovationen zu nutzen.

Unternehmen sollten fortschrittliche KI-Systeme nicht auf eine Weise entwickeln oder einsetzen, die die Demokratie untergräbt, für Einzelpersonen oder Gemeinschaften besonders schädlich sind, den Terrorismus begünstigen, kriminellen Missbrauch ermöglichen, oder erhebliche Risiken für die Sicherheit und die Menschenrechte darstellen. Derartige Bestrebungen sind inakzeptabel.

Staaten müssen ihren Verpflichtungen aus den internationalen Menschenrechtsnormen nachkommen, um sicherzustellen, dass die Rechte uneingeschränkt geachtet und geschützt werden, während die Tätigkeiten des Privatsektors im Einklang mit den internationalen Rahmenwerken, wie den Leitprinzipien der Vereinten Nationen für Wirtschaft und die OECD-Leitsätze für multinationale Unternehmen stehen müssen.

Insbesondere fordern wir die Organisationen auf, sich an die folgenden Maßnahmen zu halten, und zwar in einer Weise, die den Risiken entspricht:

1 Maßnahmen während der Einführung

Während der gesamten Entwicklung fortschrittlicher KI-Systeme sind geeignete Maßnahmen zu ergreifen, um Risiken zu identifizieren, zu bewerten und zu mindern. Dies schließt den Zeitraum vor und während ihrer Einführung und ihres Inverkehrbringens, sowie über den gesamten KI-Lebenszyklus hinweg ein.

Dazu gehört der Einsatz vielfältiger interner und unabhängiger externer Prüfmaßnahmen durch eine Kombination von Methoden zur Evaluation von Risiken und Schwachstellen, wie z.B. aktive Sicherheitsprüfungen (Red-Teaming), und die Implementierung geeigneter Maßnahmen zur Risikominderung zur Bewältigung identifizierter Risiken und Schwachstellen. Diese Test- und Abhilfemaßnahmen sollen die Vertrauenswürdigkeit und Sicherheit von Systemen in ihrem gesamten Lebenszyklus sicherstellen, so dass sie keine unangemessenen Risiken darstellen. Zur Unterstützung solcher Tests sollten Entwickler die Rückverfolgbarkeit in Bezug auf Datensätze, Prozesse und Entscheidungen während der Systementwicklung ermöglichen. Diese Maßnahmen sollten durch regelmäßig aktualisierte technische Unterlagen dokumentiert werden.

Diese Tests, um Risiken und Schwachstellen aufzudecken, sollten in sicherer Umgebung stattfinden und an mehreren Kontrollpunkten während des gesamten KI-Lebenszyklus durchgeführt werden, aber insbesondere vor der Einführung und dem Inverkehrbringen. Deren Ergebnisse dienen als Grundlage für Maßnahmen zur Bewältigung der ermittelten KI-Risiken, sowohl für die Gesellschaft, als auch in anderer Form. Unabhängig davon, ob sie versehentlich oder vorsätzlich verursacht wurden.

Bei der Entwicklung und Anwendung Ihrer Tests verpflichten sich Organisationen den folgenden Risiken gegebenenfalls Aufmerksamkeit zu widmen:

> Chemische, biologische, radiologische und nukleare Risiken, z. B. die Art und Weise, wie fortschrittliche KI-Systeme Zugriffsbeschränkungen reduzieren. Dies ist

insbesondere auch von nichtstaatliche Akteuren, für die Entwicklung von Waffen, deren Erwerb oder Nutzung besonders zu beachten.

> Offensive Cyber-Fähigkeiten, wie z. B. die Art und Weise, wie Systeme die Entdeckung von Schwachstellen ermöglichen können, sowie deren operative Nutzung. Wobei zu berücksichtigen ist, dass solche Fähigkeiten auch defensive Anwendungen ermöglichen und für die Aufnahme in ein System geeignet sein könnten.

> Risiken für Gesundheit und/oder Sicherheit, einschließlich der Auswirkungen der Systeminteraktion und der Verwendung von Werkzeugen, zum Beispiel die Fähigkeit, physische Systeme zu kontrollieren und in kritische Infrastrukturen einzugreifen.

> Risiken durch KI-Modelle, die Kopien von sich selbst erstellen oder andere Modelle "selbstreplizieren" oder trainieren.

> Gesellschaftliche Risiken sowie Risiken für Einzelpersonen und Gemeinschaften, wie z. B. die Art und Weise, wie KI-Systeme oder -Modelle schädliche Verzerrungen von Fakten vornehmen könnten und zu Diskriminierungen führen, oder für Verstöße gegen geltende Rechtnormen genutzt werden könnten, auch in Bezug auf den Schutz der Privatsphäre und den Datenschutz.

> Bedrohungen demokratischer Werte und der Menschenrechte, einschließlich der Erleichterung von Desinformation oder Beeinträchtigung der Privatsphäre.

> Risiko, dass ein bestimmtes Ereignis zu einer Kettenreaktion mit erheblichen negativen Auswirkungen führen könnte, mit Folgen für/ bis zu einer ganzen Stadt, einer ganzen Organisation oder einer ganzen Gemeinde.

Organisationen verpflichten sich, mit relevanten Akteuren aus allen Sektoren zusammenzuarbeiten, um Maßnahmen zur Risikominderung zu ergreifen, um diesen Risiken, insbesondere systemischen Risiken, zu begegnen.

Organisationen, die diese Verpflichtungen eingehen, sollten sich auch darum bemühen, die Forschung und Investitionen in die Bereiche Sicherheit, Voreingenommenheit und Desinformation, Fairness, Verständlichkeit und Interpretierbarkeit und Transparenz fortschrittlicher KI-Systeme aktiv zu fördern, sowie auf die Erhöhung der Robustheit und Vertrauenswürdigkeit fortschrittlicher KI-Systeme gegen Missbrauch hinzuarbeiten.

2 Ermittlung von Schwachstellen

Ermittlung und Minderung von Schwachstellen sowie gegebenenfalls von Sicherheitsvorfällen und Fällen von Missbrauch nach dem Einsatz einschließlich des Inverkehrbringens und Ergreifen geeigneter Maßnahmen um diese abzustellen.

Nach der Bereitstellung von neuen KI-Systemen sollten Unternehmen, soweit und wann immer dies dem Risikoniveau angemessen ist, die Systeme auf Schwachstellen, Sicherheitsvorfälle, neu auftretende Risiken und Missbrauch überwachen und geeignete Maßnahmen ergreifen, um diese zu beheben.

Organisationen werden ermutigt, beispielsweise Folgendes zu berücksichtigen:

- Für die Erkennung und Meldung von Problemen und Schwachstellen können auch die Erkenntnisse Dritter und Nutzer herangezogen werden. So könnten z. B. durch Prämiensysteme, Wettbewerbe oder Preise Anreize für die Offenlegung von Schwachstellen geschaffen werden.

Organisationen werden ferner aufgefordert, angemessene Dokumentation der gemeldeten Sicherheitsvorfälle vorzunehmen und zur Minderung der festgestellten Risiken und Schwachstellen ggf. auch die Zusammenarbeit mit anderen Interessengruppen zu suchen. Mechanismen zur Meldung von Schwachstellen, sollte für eine Vielzahl von Interessenträgern zugänglich sein.

3 Berichterstattung

Öffentliche Berichterstattung über die Fähigkeiten, Grenzen und Bereiche geeigneter und unsachgemäße Nutzung von KI-Systemen, und um ausreichende Transparenz zu gewährleisten, zu unterstützen und somit die Rechenschaftspflicht zu erhöhen.

Hierzu sollte auch die Veröffentlichung von Transparenzberichten gehören, die aussagekräftige Informationen für alle neuen und bedeutenden Veröffentlichungen im Bereich fortschrittlicher KI-Systeme enthalten.

Diese Berichte, Gebrauchsanweisungen und einschlägigen technischen Dokumentationen sollten soweit erforderlich, auf dem neuesten Stand gehalten werden und beispielsweise Folgendes enthalten:

> Einzelheiten zu den durchgeführten Evaluierungen zu potenziellen Sicherheitsrisiken, gesellschaftlichen Risiken, sowie zu Risiken für die Menschenrechte.

> Kapazitäten eines Modells/Systems und signifikante Leistungseinschränkungen, die Auswirkungen auf die Bereiche der sachgemäßen Nutzung haben.

> Erörterung und Bewertung der Auswirkungen und Risiken des Modells/ Systems für die gesellschaftliche Sicherheit, sowie in Bezug auf Diskriminierung, Gefährdung des Schutzes der Privatsphäre oder personenbezogener Daten und Fairness.

> Die Ergebnisse der Sicherheitsüberprüfungen (Red-Teaming) um die Einsatzbereitschaft des Modells/Systems über die Entwicklungsphase hinaus zu beurteilen.

Die Informationen in den Transparenzberichten sollten hinreichend klar und verständlich sein, um den Betreibern und Nutzern die Möglichkeit zu geben, die Ausgabe des Modells/Systems zu verbessern und den Nutzern eine angemessene Nutzung zu ermöglichen. Die Transparenzberichterstattung sollten zudem durch solide Dokumentationsprozesse, wie technische Dokumentation und Gebrauchsanweisungen ergänzt werden.

4 Informationsaustausch

Organisationen, die fortschrittliche KI-Systeme entwickeln, sollten auf einen verantwortungsvollen Informationsaustausch und die Meldung von Vorfällen hinarbeiten, ggf. auch unter Einbeziehung der Industrie, Regierungen, Gesellschaft und Wissenschaft.

Dazu gehört auch der verantwortungsvolle Austausch von Informationen, soweit angemessen, einschließlich, aber nicht beschränkt auf Evaluierungsberichte, Informationen über Sicherheitsrisiken, gefährliche beabsichtigte oder unbeabsichtigte Fähigkeiten und Versuche von KI-Akteuren, Schutzmaßnahmen im gesamten KI-Lebenszyklus zu umgehen.

Organisationen sollten an der Entwicklung, Weiterentwicklung, Einrichtung und Übernahme von Mechanismen arbeiten oder sich zusammenschließen, um angemessene, gemeinsame Standards, Instrumente und Mechanismen zur Gewährleistung der Sicherheit und Vertrauenswürdigkeit fortschrittlicher KI-Systeme zu schaffen.

Dazu sollte über den gesamten Lebenszyklus der KI auch eine angemessene und transparente Dokumentation darüber erfolgen, welche erhebliche Risiken die KI für die allgemeine Sicherheit und die Gesellschaft darstellen kann.

Unternehmen sollten mit anderen Organisationen über den gesamten KI-Lebenszyklus hinweg zusammenarbeiten, um Berichte zu erstellen, auszutauschen und der Öffentlichkeit relevante Informationen zur Förderung der Sicherheit und Vertrauenswürdigkeit fortschrittliche KI-Systeme zukommen zu lassen, und hierüber ggf. auch die zuständigen Behörden zu informieren.

Durch diese Berichterstattung sollten die Rechte des geistigen Eigentums geschützt werden.

5 Entwicklung, Implementierung und Offenlegung

Entwicklung, Implementierung und Offenlegung von KI-Governance- und Risikomanagement-Richtlinien, basierend auf einem risikobasierten Ansatz – einschließlich Datenschutzrichtlinien und Maßnahmen zur Risikominderung.

Organisationen sollten geeignete organisatorische Mechanismen ergreifen, um die Umsetzung von Risikomanagement- und Firmen-Richtlinien, einschließlich z. B. Rechenschaftspflicht und Verfahren zur Ermittlung, Bewertung, Vermeidung und Bewältigung von Risiken, während des gesamten Lebenszyklus der KI zu gewährleisten.

Dazu gehört auch die Einhaltung von Datenschutzrichtlinien für personenbezogene Daten, Benutzeraufforderungen und fortschrittliche KI-Systemausgaben. Von Unternehmen wird erwartet, dass sie ihre etablierten Firmen-Richtlinien und organisatorische Maßnahmen zur Umsetzung dieser Richtlinien in Bezug auf die Entwicklung und Nutzung von KI in Übereinstimmung mit einem risikobasierten Ansatz offenlegen. Dies sollte Rechenschaftspflicht und Firmen-Prozesse zur Evaluierung und Minimieren Sie Risiken, wo immer dies während des gesamten KI-Lebenszyklus möglich ist.

Die Risikomanagementstrategien sollten im Einklang mit einem risikobasierten Ansatz und über den gesamten KI-Lebenszyklus erfolgen, und die ganze Bandbreite der mit KI-Systemen verbundenen Risiken und die Strategien abdecken und ggf. regelmäßig aktualisiert werden.

Unternehmen sollten Richtlinien, Verfahren und Schulungen einführen, um sicherzustellen, dass die Mitarbeiter mit ihren Pflichten und den Risikomanagementpraktiken der Organisation vertraut sind.

6 Sicherheitskontrollen

Investieren Sie in robuste Sicherheitskontrollen und deren Einführung, einschließlich physischer Sicherheit, Cybersicherheit und Schutz vor Insider-Bedrohungen während des gesamten KI-Lebenszyklus.

Dazu kann das Sichern von Modellgewichtungen und Algorithmen, Servern und Datensätzen gehören, z. B. durch Betriebliche Sicherheitsmaßnahmen für die Informationssicherheit und die Steuerung angemessener Zugriffsrechte, sowie physische Zugangskontrollen.

Dazu gehört auch die Bewertung von Cybersicherheitsrisiken und die Implementierung von Cybersicherheit und angemessenen technischen- und organisatorischen Maßnahmen um sicherzustellen, dass die Cybersicherheit im Bezug auf fortschrittliche KI-Systeme den jeweiligen Umständen und den damit verbundenen Risiken angemessen ist.

Organisationen sollten auch über Maßnahmen verfügen, die das Speichern und Arbeiten mit den Datenkapazitäten fortschrittlicher KI-Systeme in einer angemessen sicheren Umgebung und mit begrenztem Zugang ermöglich. Damit reduzieren Sie sowohl das Risiko einer unbefugten Freigabe als auch das Risiko eines unbefugten Zugriffs. Dazu gehört die Verpflichtung, einen Prozess für das Schwachstellenmanagement einzurichten und die Sicherheit regelmäßig zu überprüfen. Maßnahmen, um sicherzustellen, dass sie auf einem hohen Standard gehalten werden und weiterhin geeignet sind den sich stätig verändernden Risiken zu begegnen.

Dazu gehört auch die Einrichtung eines robusten Programms zur Erkennung von Insider-Bedrohungen, das den Schutz ihres wertvollsten geistigen Eigentums und ihrer Geschäftsgeheimnisse gewährleistet. z. B. durch Beschränkung des Zugriffs auf unveröffentlichte Modell-Parameter.

7 Authentifizierung und Herkunft von Inhalten

Entwicklung und Bereitstellung zuverlässiger Mechanismen zur Authentifizierung und Herkunft von Inhalten, unter Nutzung der technischen Möglichkeiten, wie z. B. Wasserzeichen oder andere Techniken, die es den Benutzern ermöglichen, KI-generierte Inhalte zu erkennen.

Dazu gehört, soweit angemessen und technisch machbar, die Authentifizierung von Inhalten die mit dem fortschrittlichen KI-System eines Unternehmens erstellt wurden, und Mechanismen zur Feststellung der Herkunft der Inhalte. Die Informationen zur Daten-Herkunft (provenance data) sollte eine Bezeichnung des Diensts oder Modells

enthalten, der den Inhalt erstellt hat, muss aber keine Benutzerinformationen enthalten. Unternehmen sollten sich auch darum bemühen, Tools oder APIs zu entwickeln, die es den Benutzern ermöglichen festzustellen, ob bestimmte Inhalte mit ihrem fortschrittlichen KI-System erstellt wurden. z. B. über Wasserzeichen. Organisationen sollten zusammenarbeiten und gegebenenfalls in die Forschung investieren, um Fortschritte zu erzielen.

Organisationen werden ferner ermutigt, andere Mechanismen wie Kennzeichnungen oder Haftungsausschlüsse zu verwenden, um Nutzer, soweit möglich und angemessen, darüber aufzuklären, dass Sie mit einem KI-System interagieren.

8 Forschung

Priorisierung der Forschung, um Risiken für die Gesellschaft und allgemeine Sicherheitsrisiken zu minimieren und wirksame Maßnahmen zur Eindämmung des Klimawandels zu fördern.

Dazu gehören die gemeinsame Durchführung, Förderung der Zusammenarbeit und Investition in Forschung, um Fortschritte in Sicherheitsfragen zu erzielen und das Vertrauens in KI zu steigern, sowie die Bewältigung der größten Risiken voranzutreiben, und ferner Investitionen in die Entwicklung geeignete Instrumente zur Risikominderung.

Organisationen verpflichten sich, Forschung durchzuführen, zusammenzuarbeiten und in diese zu investieren, um eine Verbesserung der KI-Sicherheit, der Vertrauenswürdigkeit von KI und der Bewältigung wichtiger Risiken, wie z. B. die Priorisierung Forschung zur Wahrung demokratischer Werte, zur Achtung der Menschenrechte, zum Schutz von Kindern und Schutz der Rechte des geistigen Eigentums und der Privatsphäre sowie Vermeidung schädlicher Vorurteile, Fehl- und Desinformation sowie Informationsmanipulation zu erreichen. Unternehmen verpflichten sich auch, in Entwicklung geeigneter Instrumente zur Risikominderung zu investieren, sowie deren Auswirkungen im Hinblick auf Umwelt- und Klimaauswirkungen zu bedenken und deren Vorteile voll auszuschöpfen.

Organisationen werden ermutigt, Forschungsergebnisse und „Best Practices“ zur Risikominderung auszutauschen.

9 Entwicklung fortschrittlicher KI-Systeme

Priorisierung der Entwicklung fortschrittlicher KI-Systeme, um den weltweit größten Herausforderungen, insbesondere, aber nicht beschränkt auf, die Klimakrise, die globale Gesundheit und die Bildung zu begegnen.

Diese Anstrengungen werden unternommen, um die Fortschritte bei der nachhaltigen Entwicklung der Vereinten Nationen zu unterstützen, Ziele zu verfolgen und die KI-Entwicklung zum globalen Nutzen zu fördern.

Unternehmen sollten einem verantwortungsvollen Umgang mit vertrauenswürdiger und menschenzentrierter KI Vorrang einräumen und auch Initiativen zur Förderung der

digitalen Kompetenz unterstützen, die die allgemeine und berufliche Bildung der Öffentlichkeit fördern, Studierende und Arbeitnehmer einzubeziehen, damit sie vom Einsatz fortschrittlicher KI-Systeme profitieren können, und Einzelpersonen und Gemeinschaften dabei helfen, die Art, die Fähigkeiten, die Grenzen und die Auswirkungen dieser Technologien zu erfassen. Organisationen sollten mit der Zivilgesellschaft und kommunalen Gruppen zusammenarbeiten, um prioritäre Herausforderungen zu ermitteln und innovative Lösungen zu entwickeln, um diesen zu begegnen.

10 Förderung der Entwicklung und gegebenenfalls der Übernahme internationaler Technische Normen

Organisationen werden ermutigt, zur Entwicklung und gegebenenfalls zur Nutzung internationalen technischen Standards und Best Practices beizutragen. Dies betrifft die Zusammenarbeit mit Standards Development Organizations (SDOs), sowohl bei der Entwicklung von Mechanismen, Kennzeichnung der Authentifizierung und Herkunft von Inhalten, Cybersicherheitsstrategien, Berichterstattung und andere Maßnahmen. Insbesondere werden auch Organisationen ermutigt, an der Entwicklung Interoperable internationale technische Standards und Rahmenwerke, um den Nutzern bei der Unterscheidung von Inhalten zu helfen die von KI aus nicht KI-generierten Inhalten generiert werden.

11 Maßnahmen zur Dateneingabe und zum Schutz personenbezogener Daten

Unternehmen sind angehalten geeignete Maßnahmen zu ergreifen, um die Datenqualität zu gewährleisten und den Schutz personenbezogener Daten sicherzustellen, sowie geistiges Eigentum vor schädlichen Veränderungen zu schützen.

Geeignete Maßnahmen könnten Transparenz-Schulungen, Schulungstechniken zum Schutz der Privatsphäre und/oder Tests und Feinabstimmungen, um sicherzustellen, dass die Systeme keine vertraulichen oder sensiblen Daten preisgeben, enthalten.

Organisationen werden ermutigt, geeignete Schutzvorkehrungen zu treffen und die Rechte in Bezug auf Privatsphäre und geistiges Eigentum, einschließlich urheberrechtlich geschützter Inhalte zu schützen.

Organisationen sollten auch die geltenden rechtlichen Rahmenbedingungen einhalten