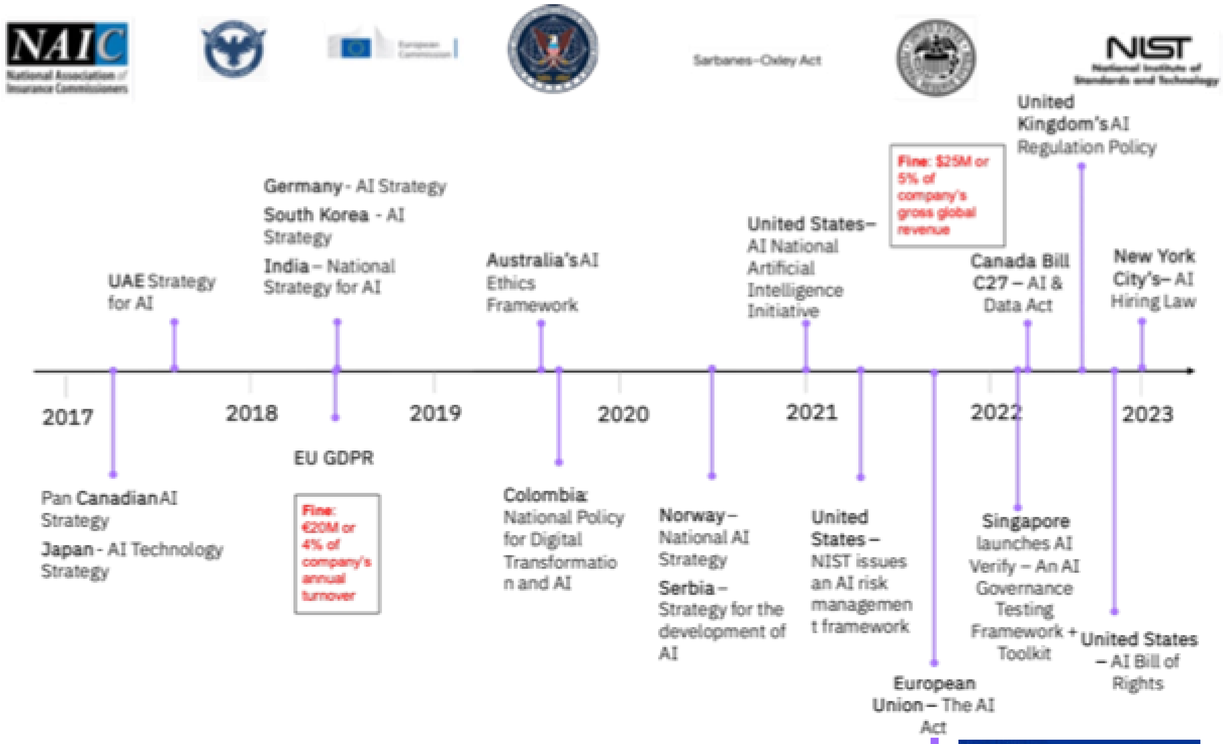
Die neue KI-Verordnung der EU: Ist KI jetzt verboten?

Anwendungsfälle nach Risikoklassen in der Praxis

Gabriele Hantschel
Manager Sustainability Software
Technology Expert Labs

Gabriele.Hantschel@de.ibm.com
https://www.linkedin.com/in/hantschelgabriele/

IBM



EU AI Act Fine: €35M or 7% of company's global revenue

1.8.2024

heise online | heise + | 5 Monate nur 5 € monatlich

IT Wissen Mobiles Security Developer Entertainment Netzp

TOPTHEMEN:

KÜNSTLICHE INTELLIGENZ | WINDOWS | ENERGIE | SMART HOME | DAT

ASTRONOMIE | PODCASTS | DOWNLOADS

heise online › Künstliche Intelligenz › "Schlechtester Paketdienst": DPD-Chatbot flucht und bes

"Schlechtester Paketdienst": DPD-Chatbot flucht und beschimpft eigene Firma

Ein Kunde bringt das KI-System von DPD dazu, allen trainierten Anstand zu vergessen und ein Haiku über die Nutzlosigkeit des Zustellers zu dichten.

Lesezeit: 3 Min. | In Pocket speichern | 218

ChatGPT Just Passed an MBA-Level Exam at Wharton

Wharton Professor Christian Terwiesch found that ChatGPT was great at open-ended questions, but struggled with some 6th grade-level math.

By Kevin Hurler | Published Tuesday 9:40AM | Comments (12) | Alerts

Terwiesch sees the promise of ChatGPT as an educational tool, but plans to ban it from his classes moving forward.

THE WALL STREET JOURNAL.

Home World U.S. Politics Economy Business Tech Markets Opinion Books & Arts Real Estate Life & W

TECH | PERSONAL TECH | PERSONAL TECHNOLOGY: JOANNA STERN

ChatGPT Wrote My AP English Essay—and I Passed

Our columnist went back to high school, this time bringing an AI chatbot to complete her assignments

The Register

Microsoft, OpenAI sued for $3B after allegedly trampling privacy with ChatGPT

HALLUZINATION

ChatGPT erfindet Gerichtsakten

Ein Anwalt wollte sich von ChatGPT bei der Recherche unterstützen lassen – das Ergebnis ist eine Blamage.

Air Canada Held Responsible for Chatbot's Hallucinations

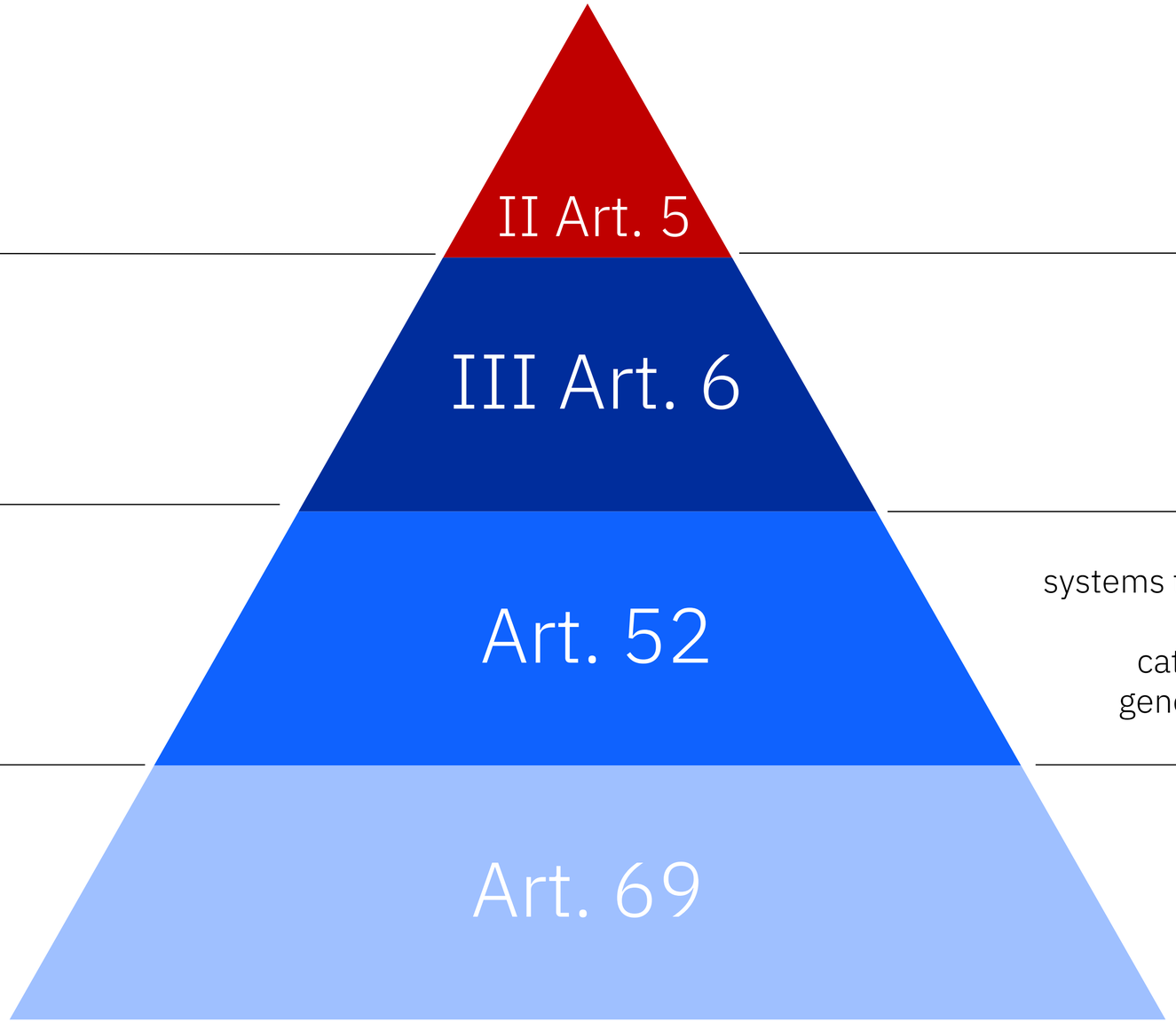
Air Canada's chatbot gave a traveler wrong airfare information. The traveler sued when the airline refused to give a refund

1 Regulatory Compliance

2 Quality

A risk based approach to regulations

Start: 1.8.2024



Prohibited

Unacceptable risk
harmful AI practices considered to be a clear threat to people's safety, livelihoods and rights

Regulated

High risk
AI systems that create adverse impact on people's safety or their fundamental rights

Transparency

Limited risk
systems that interacts with humans (i.e. chatbots), emotion recognition systems, biometric categorization systems, and AI systems that generate or manipulate image, audio or video content (i.e. deepfakes)

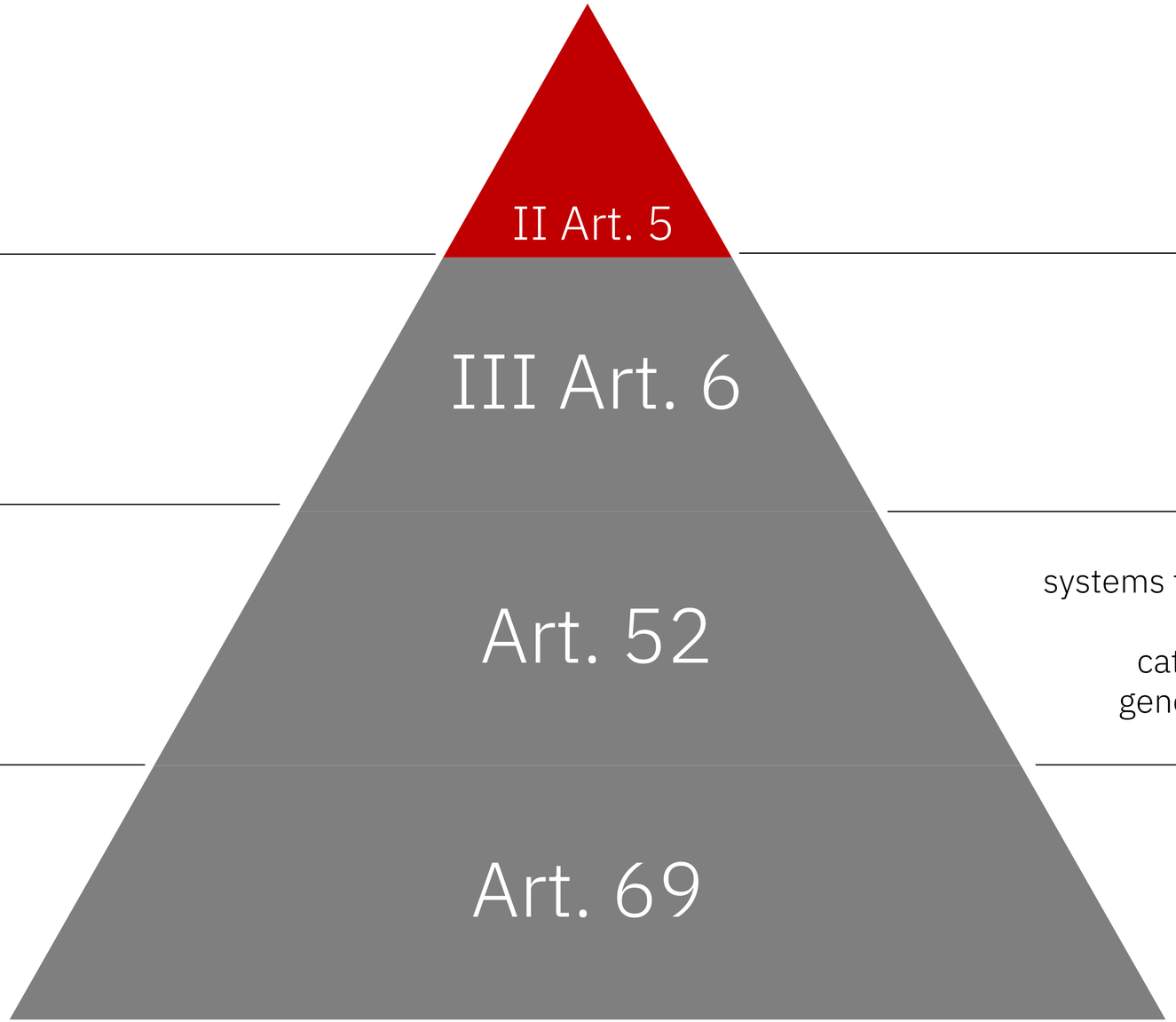
Voluntary code of conduct

Low and minimal risk
All other AI systems presenting only low or minimal risk

Conformity-Tool des Europäischen Instituts für Innovation und Technologie (EIT): EU AI Act Conformity Tool zur Auswahl der Risikoklasse für ein KI-System

A risk based approach to regulations

2/2025



Prohibited

Unacceptable risk
harmful AI practices considered to be a clear threat to people's safety, livelihoods and rights

Regulated

High risk
AI systems that create adverse impact on people's safety or their fundamental rights

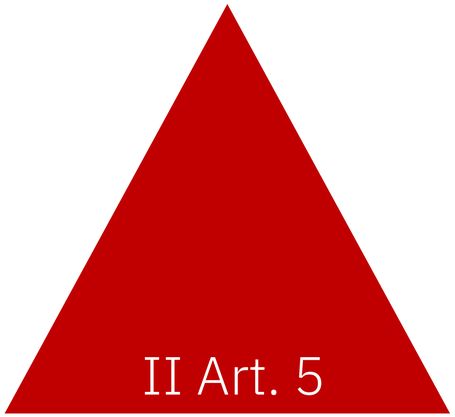
Transparency

Limited risk
systems that interacts with humans (i.e. chatbots), emotion recognition systems, biometric categorization systems, and AI systems that generate or manipulate image, audio or video content (i.e. deepfakes)

Voluntary code of conduct

Low and minimal risk
All other AI systems presenting only low or minimal risk

Source: https://www.computerbild.de/artikel/cb-News-Internet-Social-Scoring-23171771.html



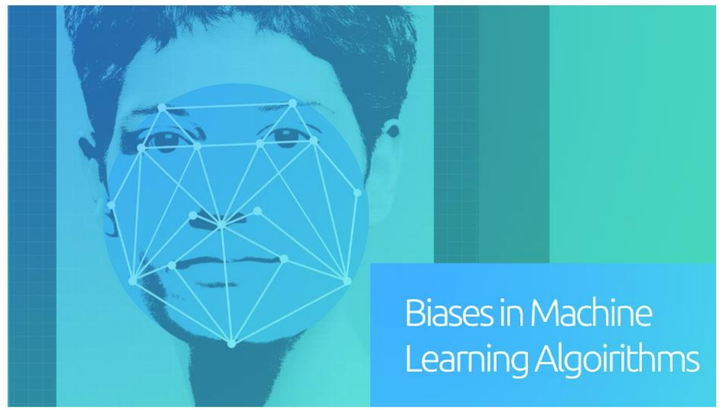
UNVERTRETBARES RISIKO: VERBOTEN

Erhebliche Risiken für Sicherheit oder Grundrechte der Menschen:

- Soziale Bewertungssysteme (**Social Scoring**) durch Regierungen wie in China,
- Manipulative KI, die das Verhalten von Menschen auf unethische Weise beeinflusst, zur Umgehung des freien Willens,
- **Echtzeit-basierte biometrische Fernidentifikation**, auch im **Nachhinein**, in öffentlichen Räumen, für Strafverfolgungszwecke,
- Biometrische **Kategorisierungssysteme**, die sensible Merkmale (Alter, Geschlecht, Ethnik, Religion, soziale Situation, etc.) verwenden,
- Vorausschauende Polizeiarbeit auf Basis von sogenanntem „**Profiling**“ (Hautfarbe, Religionszugehörigkeit, Geographie, …),
- Systeme zur **Emotionserkennung am Arbeitsplatz** und in **Bildungseinrichtungen**, ausgenommen aus medizinischen und sicherheitstechnischen Gründen,
- Beliebige **Extraktion von biometrischen Daten** aus sozialen Medien oder Video-überwachungsaufnahmen zur Erstellung von Datenbanken zur Gesichtserkennung.

Gender Bias

Carnegy Mellon University, 11.10.2018



October 11, 2018

Amazon Scraps Secret AI Recruiting Engine that Showed Biases Against Women

AI Research scientists at Amazon uncovered biases against women on their recruiting machine learning engine

Racial Prejudice

Forbes, 01.07.2015

Google Photos Tags Two African-Americans As Gorillas Through Facial Recognition Software

Maggie Zhang Forbes Staff
I write about technology, innovation, and startups.

Follow

RANCISCO, CA - MAY 28: Google Photos director Anil Sabharwal announces Google Photos
the 2015 Google I/O conference on May 28, 2015 in San Francisco, California. The annual
e I/O conference runs through May 29. (Photo by Justin Sullivan/Getty Images) [-]

Age Discrimination

Forbes, 12.11.2022

AI Ethics And AI Law Are Warning About Unchecked AI Ageism Discrimination

Lance Eliot Contributor
Dr. Lance B. Eliot is a world-renowned expert on Artificial Intelligence (AI) and Machine Learning...

Follow

Nov 12, 2022, 08:00am EST

AI ageism is a neglected and nearly unexamined discriminatory issue that requires added attention ... [+] GETTY

LGBTQAI+ Discrimination

MIT News, 11.02.2018

Study finds gender and skin-type bias in commercial artificial-intelligence systems

Examination of facial-analysis software shows error rate of 0.8 percent for light-skinned men, 34.7 percent for dark-skinned women.

Watch Video

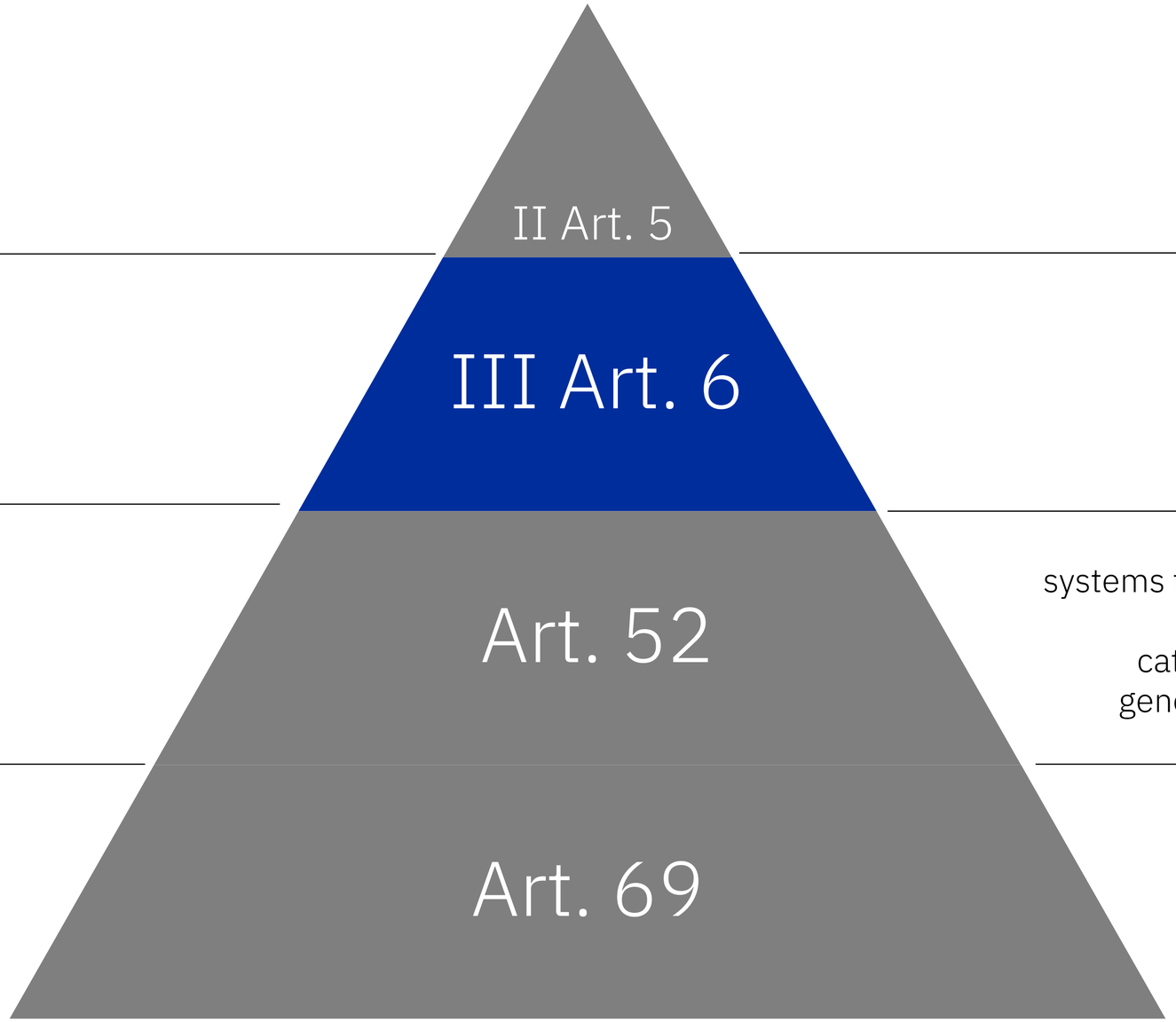
Larry Hardesty | MIT News Office

Joy Buolamwini, a researcher in the MIT Media Lab's Civic Media group

Photo: Bryce Vickmark

A risk based approach to regulations

8/2026



Prohibited

Unacceptable risk
harmful AI practices considered to be a clear threat to people's safety, livelihoods and rights

Regulated

High risk
AI systems that create adverse impact on people's safety or their fundamental rights

Transparency

Limited risk
systems that interacts with humans (i.e. chatbots), emotion recognition systems, biometric categorization systems, and AI systems that generate or manipulate image, audio or video content (i.e. deepfakes)

Voluntary code of conduct

Low and minimal risk
All other AI systems presenting only low or minimal risk

Source: https://www.computerbild.de/artikel/cb-News-Internet-Social-Scoring-23171771.html

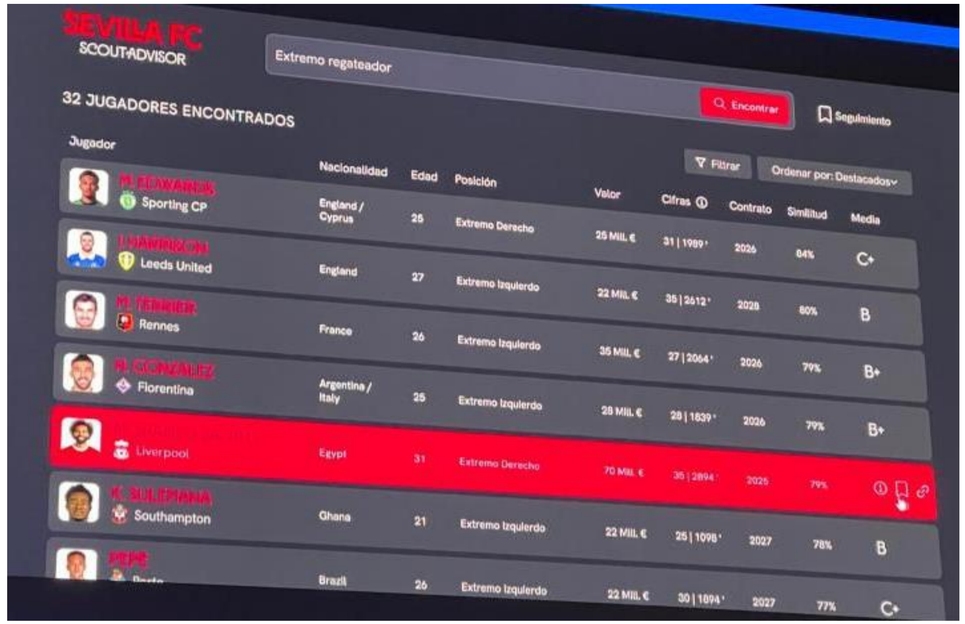
GenAI-Tool Scout Advisor: FC Sevilla sucht Spieler künftig mit IBM **watsonx**

Computerwoche, 23 Januar 2024

Um die bestmöglichen Teams zusammenzustellen, setzt der FC Sevilla beim Scouting künftig auf GenAI.
Foto: Christian Bertrand – shutterstock.com

Der **Scout Advisor** kombiniert Spielerdaten und Scouting-Berichte mit Hilfe generativer KI. Damit trifft FC Sevilla fundiertere Personalentscheidungen.

Es werden durch Eingabe von Suchbegriffen passende Spielerprofile angezeigt, mit Einbezug von Informationen aus Scouting-Berichten, die das Tool "versteht".



[AI Revolution: Sevilla FC's Recruitment Game Changer]

ISS: Roboter Cimon auf seiner ersten Mission mit Alexander …

- 2018 Start mit IBM, Airbus, DLR, ESA User Support Centrum
- Dreidimensionale schwerelose Bewegung in der ISS anhand von Sensoren und Aktoren (Airbus)
- Einbindung der Erkenntnisse der Uniklinik München zur Bewältigung psychischer Belastung in Isolation
- IBM Watson: Sprach-/ Bild-erkennung
- Watson Tone Analyzer: proaktive Analyse von Gefühlen anhand linguistischer Kriterien

2019: Auch CIMON-2 meistert seinen Einstand auf der ISS

wissenschaftliche Erforschung der Auswirkungen von Stress und Isolation bei Langzeitmissionen

III Art. 6

COMPUTERWOCHE

von Manfred Bremmer
Senior editor, IOT & Mobile, at Computerwoche.de

Scoring

Explainable AI: Bremst das Schufa-Urteil KI-Entscheidungen aus?

11 Dezember 2023 • 3 Minuten

Deutscher Bundestag

12.05.2022

Wirtschaft — Antwort — hib 232/2022

Regierung: Fünf bis 15 Prozent aller KI-Systeme ein Risiko

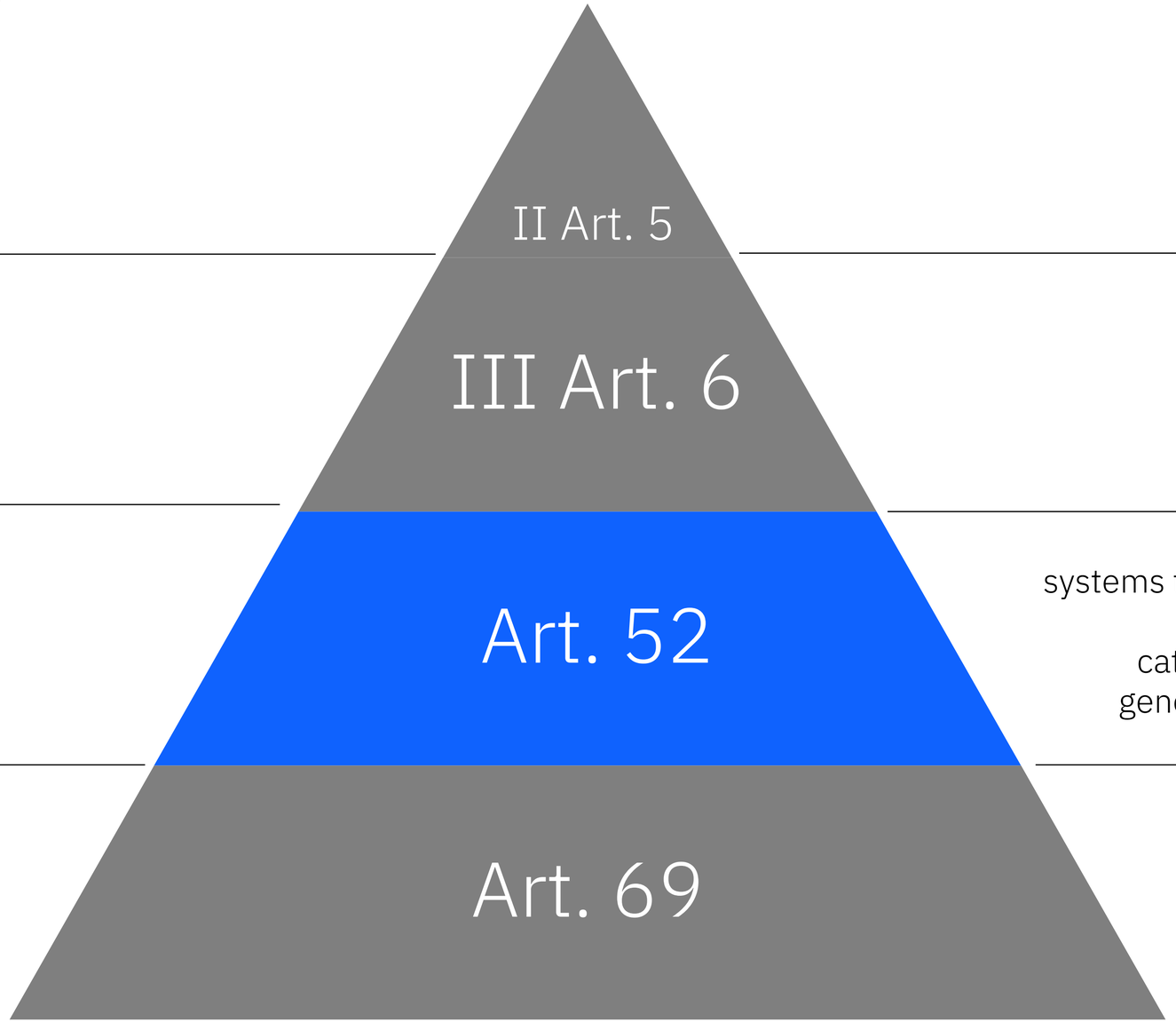
HOHES RISIKO

KI-Systeme mit Risiko für Gesundheit, Sicherheit oder Grundrechte:

- Biometrische und biometrisch-gestützte Systeme, die nicht in II, Art. 5 fallen,
- Management **kritischer Infrastrukturen** in Verwaltung und Betrieb (Energieversorgung, Verkehr, z.B. Wasser-, Gas-, Stromversorgung)
- Zugang zu wesentlichen privaten / öffentlichen Diensten: Beurteilungen für Anspruch auf öffentliche Unterstützungsleistungen, **Kreditwürdigkeit** natürlicher Personen
- **Bildung und Berufsausbildung**: Entscheidungen über Zugang zu Bildungs-einrichtungen, Bewertung von Schülern, deren Lernleistungen oder Prüfungen
- **Beschäftigung und Personalmanagement**: Auswahl von Bewerbern, Entscheidungen über Beförderungen, Kündigungen, Aufgabenzuweisungen, Leistungsüberwachung
- Priorisierung des Einsatzes von Not- und Rettungsdiensten, einschließlich Feuerwehr und **medizinischer Nothilfe**
- Anwendungen in der **Strafverfolgung** (z.B. Risikobeurteilung von Straftaten), im Asylrecht, Grenzkontrolle und der Justiz (Rechtspflege und demokratische Prozesse)
- AI Act Anhang - Produktsicherheitsregelungen: KI-Systeme als Sicherheits-komponenten in der Luftfahrt, in Spielzeug, Medizingeräten oder in.

A risk based approach to regulations

8/2026



Prohibited

Unacceptable risk
harmful AI practices considered to be a clear threat to people's safety, livelihoods and rights

Regulated

High risk
AI systems that create adverse impact on people's safety or their fundamental rights

Transparency

Limited risk
systems that interacts with humans (i.e. chatbots), emotion recognition systems, biometric categorization systems, and AI systems that generate or manipulate image, audio or video content (i.e. deepfakes)

Voluntary code of conduct

Low and minimal risk
All other AI systems presenting only low or minimal risk

Art. 52

BEGRENZTES / LIMITIERTES RISIKO

KI-Systeme, die direkt mit Menschen interagieren, müssen Transparenzpflichten erfüllen. Entwickler und Anwender müssen sicherstellen, dass die Endnutzer über die Interaktion mit KI wissen. KI-generierter Output muß als solchen deklariert werden:

- Chatbots
- Emotionserkennungssysteme
- Biometrische Kategorisierungssysteme
- KI-generierte / veränderte Audio-, Bild-, Video-,Textinhalte
- Deep Fakes, Fake News

zdfheute

Warum das Papst-Foto nicht nur witzig ist

von Julia Klaus 27.03.2023 | 16:48

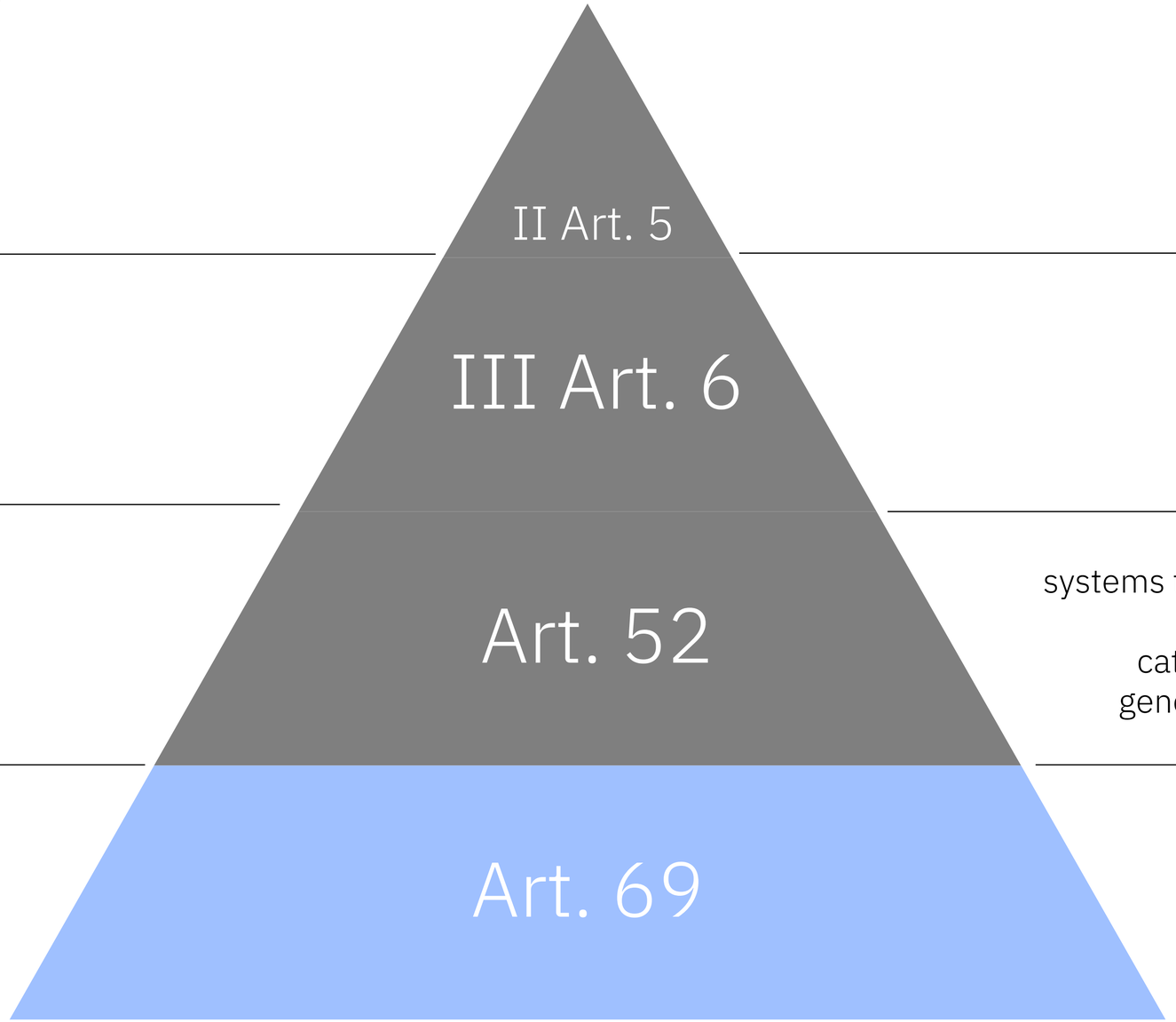
Nein, Papst Franziskus hat nicht diese hippe Daunenjacke mit dem päpstlichen Kruzifix getragen. Das Foto hat eine Künstliche Intelligenz erstellt. Warum das nicht nur witzig ist.

Lernende Systeme
GERMANY'S PLATFORM FOR ARTIFICIAL INTELLIGENCE

Quiz: Fakt oder Fake

A risk based approach to regulations

8/2026



Prohibited

Unacceptable risk
harmful AI practices considered to be a clear threat to people's safety, livelihoods and rights

Regulated

High risk
AI systems that create adverse impact on people's safety or their fundamental rights

Transparency

Limited risk
systems that interacts with humans (i.e. chatbots), emotion recognition systems, biometric categorization systems, and AI systems that generate or manipulate image, audio or video content (i.e. deepfakes)

Voluntary code of conduct

Low and minimal risk
All other AI systems presenting only low or minimal risk


SONY

Sund≋Bælt

Verlängerung der Lebensdauer um 100 Jahre
Reduzierung CO2 Emissionen um 750.000 Tonnen

“We wanted to move into the green field of digitization and go from repair and maintenance to asset management.”

Bjarne Jørgensen, Executive Director Asset Management & Operations Sund & Bælt

Solution components:
IBM® Maximo® Application Suite
IBM Maximo for Civil Infrastructure
IBM Maximo Visual Inspection

https://www.ibm.com/case-studies/sund-and-baelt

Video Sund & Bælt

autostrade per l'italia

IBM

„Mit leidenschaftlicher Arbeit wollen wir einen radikalen Wandel bei Autos erreichen: technologische Innovationen, Digitalisierung der Infrastruktur, Verbesserung der Umweltverträglichkeit und Mobilitätsdienstleistungen.“

Roberto Tomasi, Chief Executive Officer, Autostrade per l'Italia.

Solution components:
IBM® Maximo® Application Suite
IBM Maximo for Civil Infrastructure
IBM® Cloud Pak (Integration, Data)

https://www.ibm.com/case-studies/autostrade-italia/de-de/

Vorausschauende Wartung & kritische Infrastrukturen ??

Art. 69

Minimales Risiko

Diese Systeme sind weitgehend unreguliert und umfassen die meisten derzeit auf dem Markt verfügbaren KI-Anwendungen, wie z.B.:

- KI-gestützte Videospiele
- Spam-Filter
- Inventarmanagementsysteme
- Predictive Maintenance (Maschinen, Gebäude, …)
- Roboter in der Fertigung

General Purpose AI

8/2025

Gesonderte Risikokategorien

KI-Modellen mit allgemeinem Verwendungszweck (GPAI): Vielzahl von Funktionen und Anpassung an verschiedene Aufgaben

→ Risikokategorie: nicht jeweilige Anwendung, sondern Funktion des Basismodells

Für normale GPAI gilt:

- Zusätzliche technische Dokumentation
- Detaillierte Aufstellung über Verwendung urheberrechtlich geschützter Trainingsdaten
- Anforderungen zur Kennzeichnung generierter Inhalte

Für GPAI mit erheblichen Auswirkungen (systemischen Risiken):

- Pflichten in Bezug auf die Überwachung schwerwiegender Vorfälle
- Modellbewertung und Risikoevaluierung
- Angriffstests

watsonx: für IBM eigenen Modelle sind wir transparent, wie wir diese trainieren und welche Daten dazu genutzt werden: im Granite Whitepaper nachzulesen

IBM

IBM Newsroom | News | Media resources | Inside IBM | Blog

Crédit Mutuel Alliance Fédérale Accelerates Deployment of Generative AI in Collaboration with IBM

Jun 27, 2024

- Einsatz von 35 KI-Anwendungsfälle zur Unterstützung der 30.000 Bankberater bei der Kundenkommunikation (KI-Assistent mit personalisierten Antworten), Risikomanagement und Compliance
- Automatisierung der Prozesse und Abläufe mit Einsparung von 1 Mio Stunden
- Verabschiedung einer Charta für vertrauens-würdige KI mit Verpflichtungen, die den Einsatz von KI regeln (watsonx.governance).

Die eingeführte Governance

- gewährleistet die Wahrung der digitalen Privatsphäre,
- sorgt für transparente und dokumentierte KI-Nutzung,
- bevorzugt die einfachen technologischen Lösungen
- gewährleistet, dass der Grundsatz der Bündelung von Bank- und Versicherungsangeboten beibehalten wird, um die Interessen der Mitglieder und Kunden zu wahren.

Elements of AI risk

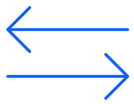
Accountability

Accuracy

Fairness

Truthfulness

Transparency



Drift

Trusted data

Sustainability

Explainability

Adversarial Robustness

IP/PII leakage

...

Regulatory Risk

Reputational Risk

Operational Risk

AI risk atlas

IBM KI Risiko Atlas

Last Updated: 2024-05-07

Explore this atlas to understand some of the risks of working with generative AI, foundation models, and machine learning models.

Risks are categorized with one of these tags:

- Traditional AI risks (applies to traditional models as well as generative AI)
- Risks amplified by generative AI (might also apply to traditional models)
- New risks specifically associated with generative AI

Risks associated with input

Training and tuning phase

Fairness

Data bias — Amplified

Robustness

Data poisoning — Traditional

Value alignment

Data curation — Amplified

Downstream retraining — New

Data laws

Data transfer — Traditional

Data usage — Traditional

Data acquisition — Amplified

Intellectual property

Data usage rights — Amplified

Confidential information in data — Traditional

Transparency

Data transparency — Amplified

Data provenance — Amplified

IBM | Documentation | Search in IBM watsonx as a Service

AI risk atlas

[IBM KI Risiko Atlas]

IBM watsonx

Change version

saas

Show full table of contents

Filter on titles

- **AI risk atlas**
  - Data bias
  - Data poisoning
  - Data curation
  - Downstream retraining
  - Data transfer
  - Data usage
  - Data acquisition
  - Data usage rights
  - Confidential information in data
  - Data transparency
  - Data provenance
  - Personal information in data
  - Reidentification
  - Data privacy rights
  - Informed consent
  - Personal information in prompt
  - Membership inference attack
  - Attribute inference attack
  - IP information in prompt
  - Confidential data in prompt
  - Evasion attack
  - Extraction attack
  - Prompt injection

Last Updated: 2024-05-07

Explore this atlas to understa

Risks are categorized with or

- Traditional AI risks (applie
- Risks amplified by generat
- New risks specifically assc

Risks associat

Training and tuning p

Docs / KI-Risiko-Atlas / Datentransparenz

Übersicht
Services und Integrationen
Einführung
KI-Risiko-Atlas
Datenverzerrung
Datenvergiftung
Datenpflege
Nachgeordnetes erneutes Training
Datenübertragung
Datennutzung
Datenbeschaffung
Datennutzungsrechte
Vertrauliche Informationen in Daten
Datentransparenz
Datenprovenienz
Personenbezogene Daten in Daten
Erneute Identifikation
Datenschutzrechte
Informierte Zustimmung
Persönliche Informationen in Eingabeaufforderung
Attacke auf Mitgliedschaftsinferenz
Attacke auf Attributinferenz
IP-Informationen in Eingabeaufforderung
Vertrauliche Daten in Eingabeaufforderung
Ausweichungsangriff
Extraktionsattacke
Injektion von Eingabeaufforderungen
Leck bei Eingabeaufforderung
Eingabeaufforderungs-Priming
Jailbreak
Ausgabeverzerrung
Entscheidungsverzerrung
Urheberrechtsverletzung

Zurück zur englischen Version der Dokumentation

Risiko der Datentransparenz für KI

Letzte Aktualisierung: 16. Aug. 2024

Risiken im Zusammenhang mit der Eingabe | Trainings-und Optimierungsphase | Transparenz | Verstärkt durch generative KI

Beschreibung

Ohne genaue Dokumentation, wie die Daten eines Modells erfasst, kuratiert und zum Trainieren eines Modells verwendet wurden, kann es schwieriger sein, das Verhalten des Modells in Bezug auf die Daten zufriedenstellend zu erklären.

Warum ist Datentransparenz für Basismodelle ein Anliegen?

Datentransparenz ist wichtig für die Einhaltung gesetzlicher Bestimmungen und die Ethik der KI. Fehlende Informationen begrenzen die Fähigkeit, Risiken im Zusammenhang mit den Daten auszuwerten. Das Fehlen standardisierter Anforderungen könnte die Offenlegung einschränken, da Unternehmen Geschäftsgeheimnisse schützen und versuchen, andere vom Kopieren ihrer Modelle abzuhalten.

Beispiel
Offenlegung von Daten-und Modellmetadaten

OpenAI's ist ein Beispiel für den Zwiespalt bei der Offenlegung von Daten und Modellmetadaten. Während viele Modellentwickler Wert darauf legen, Transparenz für Verbraucher zu ermöglichen, stellt die Offenlegung echte Sicherheitsprobleme dar und könnte die Fähigkeit erhöhen, die Modelle zu missbrauchen. Im technischen Bericht GPT-4 geben die Autoren an: "Angesichts der Wettbewerbslandschaft und der Auswirkungen großer Modelle wie GPT-4auf die Sicherheit enthält dieser Bericht keine weiteren Details zur Architektur (einschließlich Modellgröße), Hardware, Trainingsberechnung, Datasetkonstruktion, Trainingsmethode oder Ähnlichem."

Quellen: OpenAI, März 2023

Übergeordnetes Thema: AI-Risikoatlas

Wir stellen Beispiele vor, die von der Presse abgedeckt werden, um viele der Risiken der Fundamentmodelle zu erklären. Viele dieser Ereignisse, die von der Presse abgedeckt werden, entwickeln sich entweder noch weiter oder wurden gelöst, und ihre Bezugnahme kann dem Leser helfen, die potenziellen Risiken zu verstehen und auf Minderungen hinzuarbeiten. Die Hervorhebung dieser Beispiele dient nur zur Veranschaulichung.

Fairness

Data bias — Amplified

Data poisoning — Traditional

Data curation — Amplified

Downstream retraining — New

Data laws

Data transfer — Traditional

Data usage — Traditional

Data acquisition — Amplified

Intellectual property

Data usage rights — Amplified

Confidential information in data — Traditional

Transparency

Data transparency — Amplified

Data provenance — Amplified

Cookie Preferences

watsonx.governance

Accelerate responsible, transparent and explainable AI

One unified, integrated AI Governance platform to govern Generative AI and Predictive ML

https://www.ibm.com/products/watsonx-governance

Lifecycle Governance

Govern across the AI lifecycle. Automate and consolidate tools, applications and platforms. Capture metadata at each stage and support models built and deployed in 3rd party tools.

Risk Management

.

Manage risk and protect reputation by automating workflows to ensure quality and better detect bias and drift.

Regulatory Compliance

Adhere to regulatory compliance by translating growing regulations into enforceable policies.

Comprehensive
Govern the end-to-end AI lifecycle with metadata capture at each stage

Open
Support governance of models built and deployed in 3rd party tools.

Automatic metadata
and data transformation/lineage capture though Python notebooks.