

Folienzusammenstellung des Vortrags:

KI und globale Sicherheit

beim Essener Friedensforum am 15.01.2025:

<https://essener-friedensforum.de/wp-content/uploads/2024/12/242-EFF-Kuenstliche-Intelligenz.pdf>

Der Vortrag wurde mit einem speziellen Programm präsentiert, wobei zu einzelnen Folien weitere Bilder eingeblendet wurden. In dieser Zusammenstellung werden die wichtigen eingeblendeten Bilder nach der jeweiligen Folie ergänzt.

Solche Ergänzungen werden in dieser Zusammenstellung meist gekennzeichnet mit:

Bei „...“ angezeigt:

KI und globale Sicherheit



Karl Hans Bläsius, 15.01.2025

Inhalt

- KI: Einführung, Warnungen
- KI und Risiken für die globale Sicherheit
- Maßnahmen, Reduzierung Risiken

Begriff „Künstliche Intelligenz“

umstritten, viele unterschiedliche Definitionen (was ist Intelligenz?), z.B.:

- Entwicklung von Systemen, die Funktionen erfüllen, welche Intelligenz benötigen, wenn sie von Menschen ausgeführt werden.
- Automatisierung intelligenten Verhaltens
- automatisches Lösen von Problemen (Unterschied Problem – Aufgabe)
- Lösung bekannt: kein Problem mehr, keine KI ??

1956: Dartmouth-Konferenz, Workshop über 2 Monate, 10 Teilnehmer

Ansätze, Methoden der KI (Auswahl)

- logisches Schließen
- Behandlung von Unsicherheit und Vagheit
- Darstellung und Verarbeitung von Wissen
- Ontologien
- statistische Verfahren
- maschinelles Lernen
- künstliche neuronale Netze
- deep learning

Bei „logisches Schließen“ angezeigt:

Beispiel:

wenn x Kind von y ist

und y ist Kind von z ,

dann ist x Enkel von z

Bei „Behandlung von Unsicherheit und Vagheit“ angezeigt:

Unsicherheit:

wenn x **Besitzer von Auto y** , dann ist x **Nutzer von y**
(mit **Wahrscheinlichkeit w_1**)

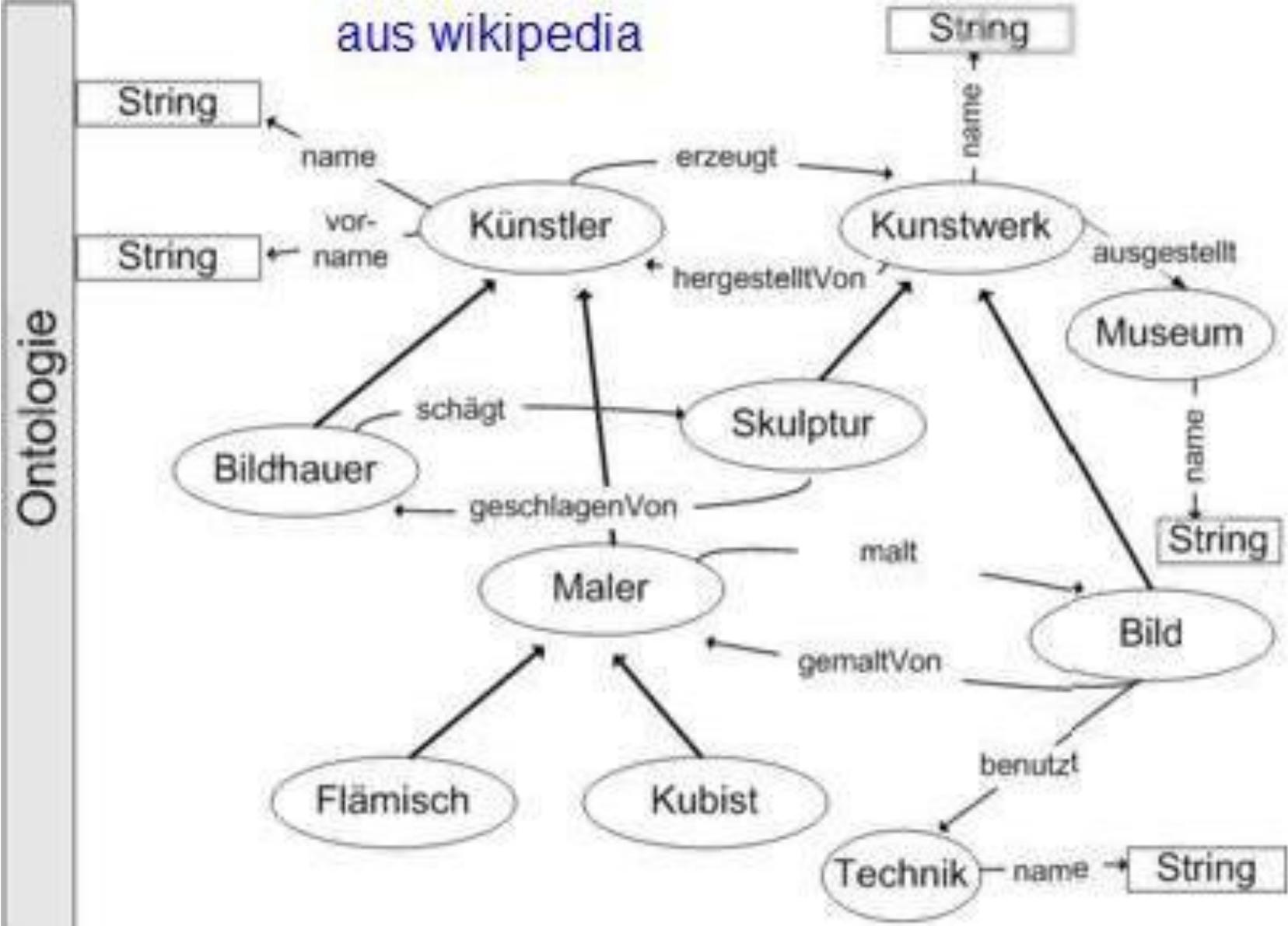
Vagheit:

Wenn x ein schweres Auto ist, benötigt x viel Kraftstoff

was bedeutet: x ist schweres Auto, x benötigt viel Kraftstoff

Behandlung von Unsicherheit, Vagheit: symbolische oder numerische Werte

Bei „Ontologien“ angezeigt :



Erkennungsaufgaben

Sehr viele KI-Anwendungen enthalten Erkennungsaufgaben

Oft prinzipiell einfach realisierbar, großer Aufwand für hohe Qualität

100% Erkennung selten realisierbar, Wahrscheinlichkeit

Erkennung nie fertig, dauerhafte Entwicklungs-, Lernaufgabe

Beispiel Klassifikation Arzt-Rechnung – Zahnarzt-Rechnung

- Ausschnitt Zahnarztrechnung
- Ausschnitt falsch klassifizierte Arztrechnung
- Fehler leicht korrigierbar, aber nicht alle Sonderfälle vorhersehbar

Bei „Ausschnitt Zahnarztrechnung“ angezeigt:

Liquidation

Für zahnärztliche Behandlung und Nebenkosten nach § 5.2 erlaube ich mir laut folgender Aufstellung zu berechnen:

EUR 102.15

Datum	Zähne	Anz.Leist.	Bezeichnung	Faktor	Betrag
18.09		1 001	Eingehende Untersuchung bei Zahn-, Mund- und Kiefererkrankungen / Parodontalbefund	2.30	12.92
	43-33	6 405	Entfernen harter und weicher Beläge am Zahn	2.30	8.40
12.10	35	1 007	Vitalitätsprüfung am Zahn	2.30	6.46
	35	1 203	Besondere Maßnahmen beim Füllen oder Präparieren	2.30	8.41
	35	1 207	Zweiflächige plastische Füllung	2.30	27.16
	35	1 209	Dreiflächige plastische Füllung	2.30	38.80

Bei „Ausschnitt falsch klassifizierte Arztrechnung“ angezeigt:

Frau
Katharina Zahn
Kirchenweg 5
123456789

BIC DAAEDED333
IBAN DE33 3006 06

Telefon: 0651
Fax: 0651
E-Mail: info@

Rechnung

Rechnungsnummer: 370

Rechnungsd

Behandelte Person: Katharina Zahn
Geburtsdatum: 01.01.1980

Sehr geehrte Frau Zahn

Generative KI – ChatGPT (Nov. 2022)

Fähigkeiten:

- Natürlichsprachliche Kommunikation
- Erzeugen von Texten, Bildern, Videos
- Programmieren

Warnungen:

- Pause KI-Entwicklung (Ende März 2023)
- Warnung vor Vernichtung durch KI (30.5.2023)



Center for
AI Safety

Statement on AI Risk

AI experts and public figures express their concern about AI risk.

Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war.

Die Minderung des Risikos des Aussterbens durch KI sollte neben anderen Risiken von gesellschaftlichem Ausmaß wie Pandemien und Atomkrieg eine globale Priorität sein.

Statement on AI Risk - Signatories

Geoffrey Hinton, Nobelpreis Physik 2024

Demis Hassabis, CEO, Google DeepMind

Sam Altman, CEO, OpenAI

Bill Gates, Gates Ventures

Ilya Sutskever, Co-Founder and Chief Scientist, OpenAI

Shane Legg, Chief AGI Scientist and Co-Founder, Google DeepMind

James Manyika, SVP, Research, Technology & Society, Google-Alphabet

Eric Horvitz, Chief Scientific Officer, Microsoft

Albert Efimov, Chief of Research, Russian Association of Artificial Intelligence

Alvin Wang Graylin, China President, HTC

Stuart Russell, Professor of Computer Science, UC Berkeley

Peter Norvig, Education Fellow, Stanford University

Generative KI - ChatGPT

Enorme Fähigkeiten (Sprache, Programmieren)

Nicht alles richtig:

- Juristische Abhandlungen (nicht exist. §)

Erwartungen:

- AGI, Superintelligenz noch dieses Jahrzehnt

Verstehen natürlicher Sprache

Beispiel:

- Wir sahen die Alpen auf dem Flug nach Süden. Wer ist geflogen?
- Wir sahen die Schwalben auf dem Flug nach Süden. Wer ist geflogen?

Allgemeines Weltwissen erforderlich

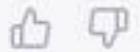
ChatGPT



Wir sahen die Alpen auf dem Flug nach Süden. Wer ist da geflogen?



Ich weiß nicht, wer auf dem Flug nach Süden geflogen ist und die Alpen gesehen hat, da keine Informationen darüber gegeben wurden.



Wir sahen die Schwalben auf dem Flug nach Süden. Wer ist da geflogen?



Es ist unwahrscheinlich, dass jemand "auf dem Flug nach Süden" Schwalben gesehen hat, da Schwalben Zugvögel sind und ihre Flüge von Norden nach Süden und umgekehrt machen. Es ist wahrscheinlicher, dass die Schwalben während eines Flugs in einer höheren Höhe gesehen wurden.



Grenzen der KI

Theoretische Grenzen:

- Nicht-Entscheidbarkeit (Prädikatenlogik 1. Stufe)
- Nicht-Berechenbare Funktionen
- Schlüsse aus unsicheren Daten sind unsicher

Solche Grenzen gelten, egal wie gut eine KI mal wird

Diese Grenzen gelten auch für Menschen

KI: viele positive Anwendungen und Perspektiven

- schwere, gefährliche Arbeiten durch Maschinen
- Gesundheit: bessere Diagnosen durch automatische Bilderkennung, basierend auf großen Datenmengen
- selbstfahrende Autos
- ...

Eigene Einstellung zu KI:

- positiv
- mögliche Gefahren sollte man kennen

Inhalt

- KI: Einführung, Warnungen
- **KI und Risiken für die globale Sicherheit**
- Maßnahmen, Reduzierung Risiken

KI und Risiken für globale Sicherheit

- Autonome Waffensysteme
- Wechselwirkungen KI – Atomwaffen, Atomkrieg aus Versehen
- Revolution der Kriegsführung durch KI
- Kognitive Kriegsführung
- Risiken durch generative KI
- Superintelligenz
- Mit KI entwickelte Biowaffen, Chemiewaffen

Autonome Waffensysteme

Autonome Waffe:

ein Gerät, das nach seiner Aktivierung mit Hilfe von Sensoren und Software selbstständig ohne menschliche Kontrolle einen Weg zu einem Ziel sucht und dort selbstständig Operationen auslösen kann, z.B. um etwas zu zerstören oder auch Menschen zu töten.

Software-Updates: zunehmende Fähigkeiten bzgl. Autonomie ?

Autonome Waffensysteme

Autonomes Fahren:

- Seit vielen Jahrzehnten möglich, bei guten Bedingungen
- Vorauss. für Freigabe: funktionieren unter allen Bedingungen

Autonome Waffensysteme:

- Vorstufen, Zwischenlösungen einsetzbar
- Kollateralschäden werden in Kriegen in Kauf genommen

Autonome Waffensysteme

Bei vielen Waffenarten anwendbar, z.B. U-Boote

Beispiel: Slaughterbots (FLI)

TAB: Eskalationspotenzial

TAB: Strategische Stabilität

Bei „Beispiel: Slaughterbots (FLI)“ wurde ein Video gezeigt. Dieses war ein Ausschnitt von:

<https://www.youtube.com/watch?v=9CO6M2HsoIA>

Bei „TAB: Eskalationspotenzial“ wurde gezeigt:

Deutscher Bundestag

19. Wahlperiode

Drucksache 19/23672

21.10.2020

Bericht

**des Ausschusses für Bildung, Forschung und
Technikfolgenabschätzung (18. Ausschuss)
gemäß § 56a der Geschäftsordnung**

Technikfolgenabschätzung (TA)

Autonome Waffensysteme

bar. So könnten das operative Geschehen und die Entscheidungsprozesse durch AWS derart beschleunigt werden, dass Menschen kognitiv und hinsichtlich ihres Reaktionsvermögens an ihre Grenzen kämen. So könnte in einer Krise eine Eskalationsspirale automatisiert und möglicherweise ungewollt in Gang gesetzt werden. Für ein solches Szenario wurde der Begriff »flash war« geprägt.

Bei „TAB: Strategische Stabilität“ wurde gezeigt:

Auf der globalen Ebene spielt das strategische Gleichgewicht zwischen den Nuklearwaffenstaaten nach wie vor eine herausragende Rolle. Es basiert wesentlich auf der gesicherten Fähigkeit eines Zweitschlags und der daraus resultierenden Abschreckung eines möglichen Erstschlags. Es wäre vorstellbar, dass sehr potente AWS zukünftig als konventionelle Erstschlagwaffen zur Zerstörung gegnerischer Nuklearwaffenarsenale eingesetzt werden könnten, die mögliche Ziele (Raketensilos oder mit Nuklearwaffen bestückte U-Boote) selbstständig aufklären, in deren Nähe unentdeckt verweilen und auf Befehl koordiniert diese Ziele angreifen und zerstören. AWS könnten auch als Trägerplattformen für Nuklearwaffen verwendet werden, beispielsweise in Form von autonomen Unterwasserfahrzeugen. Diese könnten schneller, überraschender und koordinierter als bisherige Trägersysteme zuschlagen und vorhandene Verteidigungsmaßnahmen aushebeln. Eine solche Nutzung von AWS würde die strategische Stabilität massiv infrage stellen. Dies wiederum könnte weitere nukleare Abrüstung unmöglich machen und eine Ära nuklearer Modernisierung oder gar nuklearer Aufrüstung einläuten.

Autonome Waffensysteme

Koalitionsvertrag 2021:

- „Letale Autonome Waffensysteme, die vollständig der Verfügung des Menschen entzogen sind, lehnen wir ab. Deren internationale Ächtung treiben wir aktiv voran.“

Helsing, München:

- Autonome Drohne vorgestellt (Nov. 2024)
- 4000 Drohnen für die Ukraine, von Regierung finanziert

Bei „Autonome Drohne vorgestellt (Nov. 2024)“ wurde ein Bild gezeigt von:

<https://esut.de/2024/12/fachbeitraege/55118/helsing-stellt-intelligente-kampfdrohne-hx-2-karma-offiziell-vor/>

KI und Risiken für globale Sicherheit

- Autonome Waffensysteme
- Wechselwirkungen KI – Atomwaffen, Atomkrieg aus Versehen
- Revolution der Kriegsführung durch KI
- Kognitive Kriegsführung
- Risiken durch generative KI
- Superintelligenz
- Mit KI entwickelte Biowaffen, Chemiewaffen

FWES - Frühwarn- und Entscheidungssysteme

Ziel: frühe Erkennung eines Angriffs mit Atomraketen

Aufbau:

- Sensoren (z.B. Satelliten, Radar, Horchsensoren in Meeren)
- Computerzentren und Kommunikationsnetzwerke zur Analyse und Übermittlung von Daten
- Kommandostellen zur Bewertung von Warninformationen und der Gefährdungslage sowie zur Planung und Anordnung von Aktionen.

Problem: Gefahr eines Atomkriegs aus Versehen
durch Computer-Fehler (Fehlalarm)

Risiko „Atomkrieg aus Versehen“

Fehlalarm als einzelnes Ereignis in Friedenszeiten ist unkritisch.

Es gab aber einige kritische Situationen, zum Glück ist nichts passiert.

Risiko „Atomkrieg aus Versehen“ steigt, falls:

- Krisensituation vorliegt (weltpolitische Lage, Drohungen)
- ungewöhnliche Fehler auftreten
- mehrere Ereignisse gleichzeitig zusammentreffen
- handelnde Personen Fehleinschätzungen vornehmen

Vorfälle und Alarmmeldungen

1962, Kuba-Krise: mehrere Alarmmeldungen und Vorfälle, u.a. Archipow

26. September 1983:

- **Stanislaw Petrow** verhindert Atomkrieg aus Versehen
- kritische politische Lage (u.a. Nachrüstung im Westen)
- Abschuss koreanische Passagiermaschine durch Sowjets kurz vorher
- **Petrow: Angriff mit nur 5 Raketen unwahrscheinlich**

12. Dezember 2020: Warnung vor Raketenangriff in Ramstein

Fehlalarme werden üblicherweise nicht bekannt (Geheimhaltung)

Bei „Stanislaw Petrow verhindert Atomkrieg aus Versehen“ wurde gezeigt:

RETTER DER WELT

Oberhausen erinnert als erste Stadt an den Retter der Welt **WAZ+**

Michael Bresgott

19.05.2019, 14:29

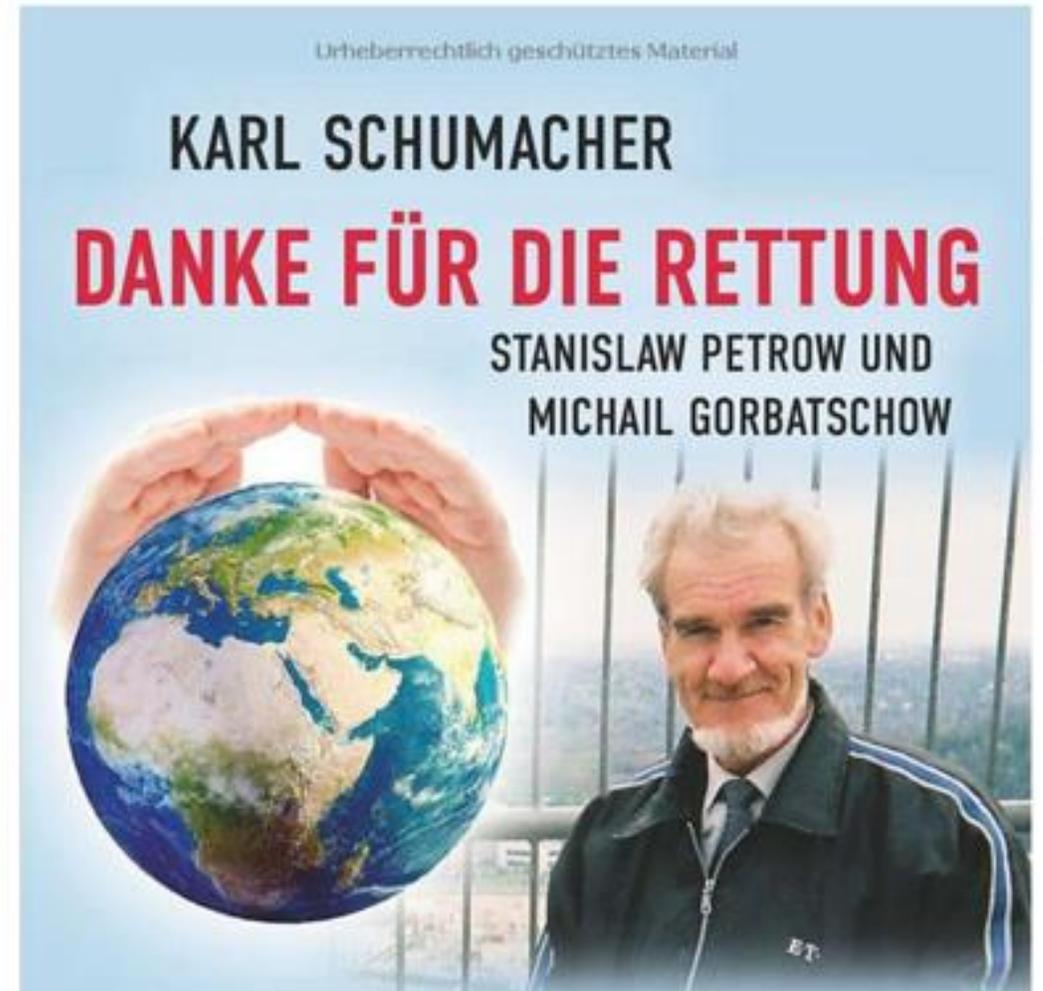
Lesedauer: 3 Minuten



Die Kinder von Stanislaw Petrow, Dmitry und Elena, am Sonntag mit Karl Schumacher an der dreisprachigen Gedenktafel für ihren im Mai 2017 verstorbenen Vater.

Foto: Kerstin Bögeholz

Auf dem Foto: Sohn und Tochter von Petrow und Karl Schumacher



Erkennung Raketenangriffe

- Erkennung unsicher
- Erkennung und Risiko hängt von vielen Zufällen ab

Fehlalarm bei Petrow:

- Anzahl 5 Raketen zufällig
- Datum zufällig

Ursache Unfälle: zufällige Ereignisse fallen zeitlich zusammen

- Alltag: Unfälle basierend auf Zufällen akzeptiert
- Atomwaffen: Unfälle basierend auf Zufällen nicht akzeptabel

Risiko „Atomkrieg aus Versehen“ steigt

Neue Technische Entwicklungen:

- Bewaffnung Weltraum
- Hyperschallraketen
- Cyberkriegskapazitäten
- KI, Autonome Waffensysteme

Folgen: Bedrohungslagen werden immer komplexer,
Alarmmeldungen immer schwerer bewertbar,
Gesamtsystem von Frühwarnkomponenten und Nuklearwaffen
immer schwerer beherrschbar

FWES – Künstliche Intelligenz (KI)

- zunehmende Anzahl von Objekten im Luftraum
- zunehmende Vielfalt an Objekttypen im Luftraum
- Vorwarnzeiten immer geringer (z.B. Hyperschallwaffen)
- Bewertung von Sensordaten immer schwieriger
- KI-Methoden zur Bewertung erforderlich

Daten- und Informationsgrundlage für KI-Entscheidungen:

Vage, unsicher, unvollständig

Unvollständigkeit – Vagheit – Unsicherheit

Frühwarnsysteme: Viele Sensoren, viele Eingangsdaten

Informationen sind

- **unvollständig**: u.a. wegen elektronischer Kampfführung
- **vage**: Größe von Radarsignal, Helligkeit von Raketensignal, ...
- **unsicher**: Erkennung Objekttyp, Objektmerkmale, ...

Annahmen sind erforderlich: können falsch sein

KI-Entscheidungen

- plausibles Schließen bei unvollständigen Informationen notwendig
- Schlüsse mit unvollständigem und unsicherem Wissen können falsch sein
- auch als sicher eingestufte KI-Entscheidungen können falsch sein
- KI: Hunderte Merkmale, oft keine einfachen Begründungen möglich
- von Menschen kaum (nicht) überprüfbar
- Problem der Unsicherheit mit KI nicht lösbar
- ➔ Wichtig für Bewertung: Erwartungshaltung, Vertrauen

Atomkriegsrisiko

Gefährliche Wechselwirkungen: KI (autonome Waffen) – Cyber – Atomwaffen
=> für Menschen unbeherrschbar (extrem komplex, geringe Entscheidungszeiten)



Vorwarnzeiten kleiner

Krisen (u.a. Klima)

Atomkrieg aus Versehen:
plötzlich, unerwartet, als Unfall

Atomkriegsrisiko – Zitat

Lahl, Varwick in „Sicherheitspolitik verstehen“ (2022):

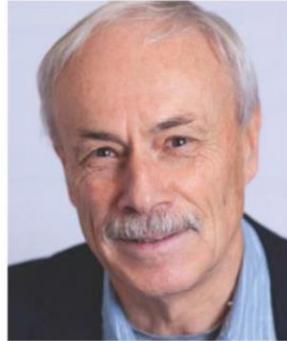
„Im Ergebnis lassen sich drei Folgerungen ableiten:

Zum **ersten** besitzen Nuklearwaffen ein Schadenspotenzial, welches das Überleben der gesamten Menschheit unter hohes Risiko stellt.

Genau deshalb konnten sie **zweitens** bisher eine eher krisenstabilisierende Rolle in der internationalen Praxis einnehmen – weil ein konkreter Einsatz für alle Seiten katastrophal wäre.

Es gibt aber **drittens** keine Garantie, dass dies immer so bleibt. Ganz im Gegenteil: Je mehr nukleare Akteure ‚mitspielen‘, je ausgereifter die technischen Entwicklungen werden und je komplexer sich damit das strategische Entscheidungsfeld um nukleare Einsätze und Einsatzdrohungen gestaltet, desto höher wird das Risiko einer mangelnden internationalen Beherrschbarkeit der Kategorie nuklearer Waffen.“

Bei „Lahl, Varwick in „Sicherheitspolitik verstehen“ (2022):“ angezeigt:



Generalleutnant a. D. Kersten Lahl war – nach einer mehr als 40-jährigen soldatischen Laufbahn – Präsident der ‚Bundesakademie für Sicherheitspolitik‘ in Berlin. Heute publiziert er zu sicherheitspolitischen Themen und wirkt im Vorstand der ‚Gesellschaft für Sicherheitspolitik‘ mit.



Prof. Dr. Johannes Varwick lehrt Internationale Beziehungen und europäische Politik an der Universität Halle-Wittenberg. Zuvor war er u. a. Leiter des Bereichs europäische Sicherheit bei der ‚Deutschen Gesellschaft für Auswärtige Politik‘ in Berlin. Zudem war er 2019–2021 Präsident der ‚Gesellschaft für Sicherheitspolitik‘.

Bildausschnitt von der Rückseite des Buches

KI und Risiken für globale Sicherheit

- Autonome Waffensysteme
- Wechselwirkungen KI – Atomwaffen, Atomkrieg aus Versehen
- Revolution der Kriegsführung durch KI
- Kognitive Kriegsführung
- Risiken durch generative KI
- Superintelligenz
- Mit KI entwickelte Biowaffen, Chemiewaffen

Revolution der Kriegsführung durch KI

OODA-Loop, Kill-chain verkürzen, KI-Einsatz in allen Phasen

OODA = observe, orient, decide, act

Beobachten, Orientieren, Entscheiden, Handeln

Clausewitz'sche Lehre: statt Verteidiger, jetzt Angreifer im Vorteil

Komplexität und Dynamik erschweren Abwehrmaßnahmen

Kriegsrisiko höher: wer zu lange wartet verliert

Kognitive Kriegsführung

Neues Gebiet der Nato zur Manipulation von Menschen

Ziele:

- Gegner beeinflussen, auch Bevölkerung
- Eigene Bevölkerung auf Kurs des Militärs bringen

Methoden:

- Microtargeting (analog Cambridge Analytica)
- usw

Bei „Neues Gebiet der Nato zur Manipulation von Menschen“ gezeigt:



Cognitive Warfare, a Battle for the Brain

Francois du Cluzel
ACT
Norfolk, Virginia
UNITED STATES OF AMERICA

Francois-pierre.ducluzelderemaurin@act.nato.int

ABSTRACT

With the growing role of technology and information overload, individual cognitive abilities will no longer be sufficient to ensure an informed and timely decision-making, leading to the new concept of Cognitive Warfare.

KI und Risiken für globale Sicherheit

- Autonome Waffensysteme
- Wechselwirkungen KI – Atomwaffen, Atomkrieg aus Versehen
- Revolution der Kriegsführung durch KI
- Kognitive Kriegsführung
- Risiken durch generative KI
- Superintelligenz
- Mit KI entwickelte Biowaffen, Chemiewaffen

ChatGPT

Anwendungen:

- Erzeugen von Texten, Bildern
- Teile verwendbar
- Gemischte Texte, teils manuell, teils automatisch
- Analog Bilder, Videos
- Evt. auch wissenschaftliche Texte mit ChatGPT-Teilen
- Nicht feststellbar was manuell, was maschinell erzeugt
- Schwer prüfbar was korrekt ist
- Quellenhinweise: teils auch automatisch erzeugt

ChatGPT und viele andere Generative KIs

- Neue Inhalte leicht erstellbar => Flut an Informationen
- Vieles falsch, Flut an Desinformationen, Deep-Fakes
- Verunsicherung, Manipulationsmöglichkeiten
- Bevölkerung: schwer feststellbar war korrekt ist
- Politik: schwer feststellbar war korrekt ist
- **Folgen für politisches Handeln ?**
- **Folgen für Demokratie ?**

ChatGPT und viele andere Generative KIs

Weiterentwicklungen, Perspektive : ?

- Kombination mit symbolischer KI
- KI - Fähigkeiten: Logisches Schließen, Wissensstrukturen, Planen
- KI besser als Menschen: Wissen, Problemlösen, Urteilsvermögen
- Bevölkerung: Orientierung an KI-Ergebnissen
- Politik: Orientierung an KI-Ergebnissen
- **Folgen für politisches Handeln ? Folgen für Demokratie ?**
- **KI entscheidet (alles) ?**

ChatGPT, Generative KIs

Zu erwarten: Cyberangriffsfähigkeiten :

- Autonome Systeme !
- Autonome Cyberwaffen !
- „Flash War“ im Internet ?
- Cyberwar zwischen verschiedenen generativen KIs ?
- Kettenreaktion von einem System durch Gewichtsänderung ?

Mögliche Folgen:

- Zusammenbruch Informationsaustausch im Internet
- Zusammenbruch Finanzwesen, Handel, ... ??

FLI, 2.4.2024, Risiken:

- explosionsartige Zunahme von Fehlinformationen und digitalen Imitationen,
- weit verbreitete Automatisierung, die Millionen von Menschen zu wirtschaftlicher Entmachtung verurteilt,
- Ermöglichung des Baus biologischer und chemischer Waffen durch Terroristen,
- extreme Konzentration von Macht in den Händen einiger weniger nicht gewählter Personen,
- vieles mehr.

Führungskräfte der KI-Konzerne

- räumen Risiken ein, machen aber mit höchster Priorität weiter
- experimentieren so lange, bis es offensichtlich gefährlich wird

Sicherheit und Verantwortung müssen von außen aufgezwungen werden.

AGI - Superintelligenz

AGI: Artificial General Intelligence:

- In vielen Bereichen ähnliches Intelligenzniveau wie Menschen

Intelligenzexplosion

Superintelligenz:

- Menschen weit überlegen
- Eventuell mehrere Superintelligenzen
(abhängig von Intelligenzexplosion und Dauer)
- Folgen: völlig unkalkulierbar

Wettrennen um Superintelligenz

Entwicklung durch

- Unternehmen (OpenAI, Google, Apple, ...)
- Staaten (China, Indien?, Russland? ...)

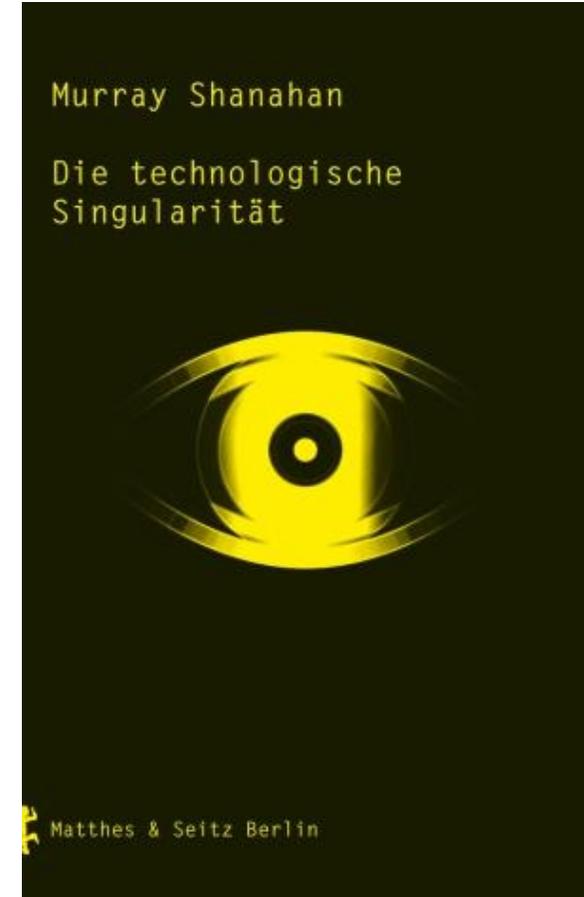
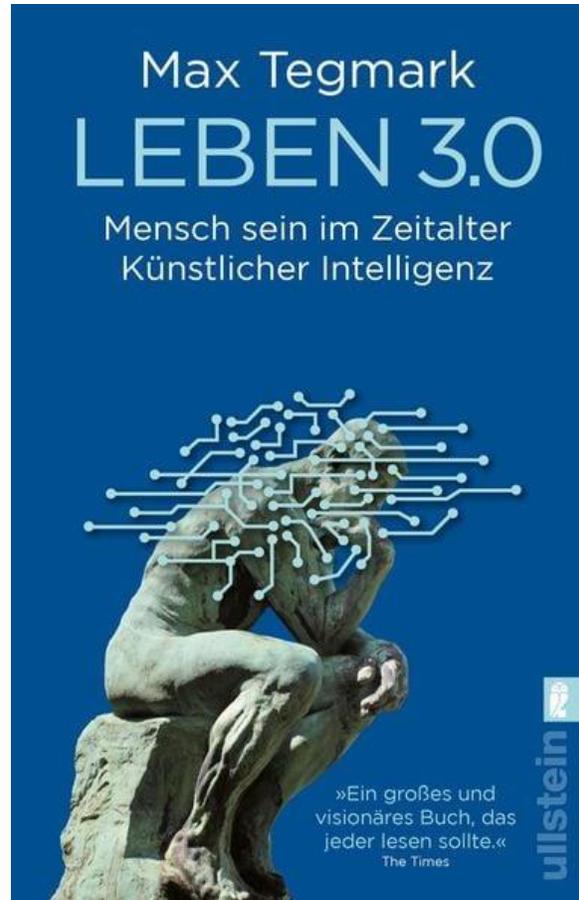
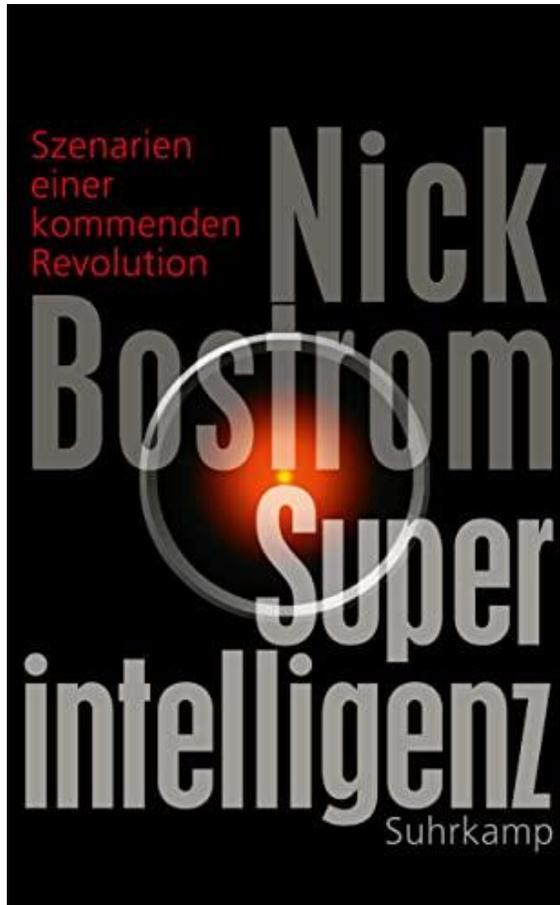
Risiken bekannt, aber mitmachen oder verlieren

Stuart Russell: Sprache und Planen reichen

Peter Norvig: AGI bereits da



Superintelligenz: wichtige Bücher



Superintelligenz

Gravierende Risiken möglich, lange bevor Superintelligenz entsteht:

- bisher unbekannte Cyberbedrohungen
- Cyberkrieg durch Systeme
- Übernahme Informationsdarstellung und -austausch
- Gesellschaften instabil

Dazu nicht erforderlich:

- Willen, Bewusstsein oder ähnliches der Maschine

KI gefährlicher als Atomwaffen ??

	Atomwaffen:	KI – SI:
Entwicklungsziel:	Waffe	Dual-Use, Positiv
Kontrolle:	Mensch	Maschine
Auswirkungen:	Lokal, evt. global	global

Warnungen vor KI: Auslöschen der Menschheit

Risiko Terminator? KI drückt Atomknopf:

- Theoretisch möglich (über Netzwerke, Cyberattacken)
- Eher unwahrscheinlich

Vergleichbares Risiko, auch ohne Superintelligenz:

- Generative KI erwirbt Fähigkeiten der kognitiven Kriegsführung
- KI manipuliert Mensch, sodass dieser Atomknopf drückt

Viele globale Risiken

- Klimawandel
- Atomwaffen, Atomkriegsrisiko
- Autonome Waffensysteme
- Mit KI entwickelte Bio- und Chemiewaffen
- Cyberwaffen, Cyberkrieg
- Weltraumbewaffnung
- Revolution der Kriegsführung durch KI
- Desinformationen, kognitive Kriegsführung
- ChatGPT: Arbeit, soziale Verwerfungen, ...
- ChatGPT: Internetdominanz, -zusammenbruch
- Kontrollverlust durch generative KI
- AGI, Superintelligenz

Inhalt

- KI: Einführung, Warnungen
- KI und Risiken für die globale Sicherheit
- **Maßnahmen, Reduzierung Risiken**

Aktuelle Forderungen

- Erhöhung Verteidigungsausgaben
- Verbesserung militärischer Stärke
- Verbesserung nukleare Abschreckung

Ziele:

- Schutz vor Aggression
(von wem? Russland ? China ? Weiteren?)

Neue Bedrohungsarten

können auch von Gruppen, Einzelpersonen, Unternehmen ausgehen

- Cyberbedrohungen
- Bedrohungen durch autonome Waffen
- KI-basierte Biowaffen, Chemiewaffen
- AGI, Superintelligenz

Atomwaffen, militärische Stärke schützen nicht vor

- Atomkrieg aus Versehen
- Internetchaos durch Verselbständigung von Systemen wie ChatGPT
- Folgen einer Superintelligenz
- Angriffe durch private Akteure oder Terroristen wie z.B.:
 - Schwerwiegende Cyberangriffe
 - Autonome Waffen (slaughterbots)
 - Mit KI erzeugten Biowaffen

Problematik technischer, militärischer Lösungen

KI-Entscheidungen evtl. besser, aber:

- Informationsgrundlage: vage, unsicher, unvollständig
- Entscheidungen: unsicher, können falsch sein
- Zeitrahmen minimal
- Menschliche Überprüfbarkeit: kaum, nicht möglich
- Eventuell totaler Kontrollverlust durch Menschen
- Gravierende Auswirkungen in sehr kurzen Zeiträumen
- Eine Eskalationsspirale beenden ist evtl. schwer möglich

Institution Krieg überwinden

Carl Friedrich von Weizsäcker-Gesellschaft
Wissen und Verantwortung e.V.

12. Februar 2023

Justus Frantz General a.D. Harald Kujat Dr. Bruno Redeker Professor Dr. Horst Teltschik

Heute ist die Situation eine andere. Und folgen wir v. Weizsäcker, so ist sie „**grundlegend anders als alle früheren.**“ Heute **nötigt** sie dazu, die Botschaft des Gleichnisses als diesseitige und dauernde Aufgabe wahrzunehmen. Der Grund liegt in der modernen Technik, „die den Krieg in eine totale Katastrophe verwandeln kann.“ „Früher haben nicht immer die Völker, aber hat doch die Menschheit die größten, damals technisch möglichen Kriege überlebt. Der Krieg war eine schreckliche, aber eine mögliche Institution. Möglich ist er noch heute, aber für sie (die Menschheit) nicht permanent überlebar; es ist notwendig, ihn als Institution zu überwinden.“

Nie zuvor war der Mensch einem Armageddon so nahe wie heute, war der Faden des nuklearen Damoklesschwertes über unseren Häuptern so dünn wie heute, das Bewusstsein der Gefahr so wenig präsent wie heute. **Und die Kuba-Krise lehrt: ein Fehler genügt!**

unter der Bedingung moderner Technik dürfen Kriege nicht mehr geführt werden.

Konfrontationskurs: Westen - Russland - China

Neuer Rüstungswettlauf auf Technologie-Feldern wie KI, Cyberraum

- Jede Nation muss vorne dabei sein
- Mit möglichen Konkurrenten mithalten können

Softwarebasierte Waffen:

- Entwicklung im Verborgenen, beliebig viele Varianten, Kopien
- Rüstungskontrolle, -vereinbarungen schwierig
- Verbreitung schwer kontrollierbar, Abrüstung schwierig
- Was jetzt an autonomen Waffen entwickelt wird, bleibt immer erhalten
- Private Akteure

Konfrontationskurs

Folgen des aktuellen Konfrontationskurses:

- gefährliche Waffenarten: Weiterentwicklungen, werden noch gefährlicher
- Superintelligenz: verstärkte Anstrengungen (China, Indien, Russland ...)
- Risiken steigen, in vielen Bereichen, auch Atomkriegsrisiko
- Politik könnte Kontrolle verlieren (KI, private Akteure in Kriegen)

Konfrontationskurs: Westen – Russland – China

Rüstungswettlauf weiter betreiben ?

- Cyberwaffen, KI, Autonome Waffen,
- Atomwaffen, Biowaffen, Weltraum, ...
- AGI, Superintelligenz

Oder einen solchen Rüstungswettlauf verhindern ?

Wie ?

Optionen Abschreckung von Cyberangriffen

Lahl, Varwick:

- Option 1: Strafe androhen
- Option 2: Sicherheit durch Resilienz
- Option 3: Internationale Verflechtungen
- Option 4: Internationale Normensetzung

Optionen auch anwendbar:

- KI, autonome Waffen, Atomkriegsrisiko
- Option 3: Prioritäten verändern sich

Zitat Lahl, Varwick:

Buch „Sicherheitspolitik verstehen“, 3. Auflage, Seite 118:

„Internationale Verflechtung. Dieser Ansatz nutzt die Erkenntnis, dass in einer global vernetzten Welt jede gewaltsame Auseinandersetzung unter dem Strich nur Verlierer hervorbringt. Je stärker also die Akteure miteinander wirtschaftlich, technologisch, kulturell und ggf. auch militärisch vernetzt sind, desto geringer ist die Chance, durch Aggression einseitige Vorteile erzielen zu können. Dieser oft zu Unrecht als naiv empfundene Gedanke läuft damit auf eine Art der Selbstabschreckung hinaus.“

„Wandel durch Handel“ : gescheitert?

„~~Wandel~~ durch Handel“ : gescheitert?

Sicherheit

„Sicherheit durch Handel“ : gescheitert?

Oder: nicht hinreichend, aber notwendig ?

Unterschied: hinreichende, notwendige Bedingung

„Sicherheit durch Handel“ : gescheitert?

Oder: nicht hinreichend, aber notwendig ?

Unterschied: hinreichende, notwendige Bedingung

Was bedeutet:

- A ist hinreichend für B ? Wenn A, dann B
- A ist notwendig für B ? Wenn B, dann A

„Sicherheit durch Handel“ : gescheitert?

Oder: nicht hinreichend, aber notwendig ?

Unterschied: hinreichende, notwendige Bedingung

Was bedeutet:

- A ist hinreichend für B ? Wenn A, dann B
- A ist notwendig für B ? Wenn B, dann A

„Sicherheit durch Handel“ nicht hinreichend, um Ukraine-Krieg zu verhindern,
aber notwendig für dauerhafte globale Sicherheit

Verhandlungen - Vereinbarungen

Wichtige Forderungen:

- Abrüstung, Rüstungskontrolle bei Atomwaffen
- Wirksame Vereinbarungen gegen Klimawandel
- Vereinbarungen zu Cyberwaffen, autonome Waffen
- Regulierung KI, AI Act reicht noch nicht

Notwendige Voraussetzungen:

- Wirksame Vereinbarungen nur mit Russland / China möglich
- Konfrontationskurs beenden (möglichst bald !!)
- Krieg beenden, Verhandlungen

Risikoreduzierung

Verbesserung Vertrauen, Kommunikation, Zusammenarbeit

zwischen allen Nationen:

- politisch, wissenschaftlich, technologisch, wirtschaftlich
- Kultur, Sport, Städtepartnerschaften, Schüleraustausch, privat

Globalisierung, wirtschaftliche Kooperation:

- Kriege eigentlich nicht mehr führbar (Selbstabschreckung)

Wichtige Forderungen:

- Abrüstung, Rüstungskontrolle bei Atomwaffen
- Vereinbarungen zu Cyberwaffen, autonome Waffen
- **Regulierung KI (weltweit, UN-Ebene)**

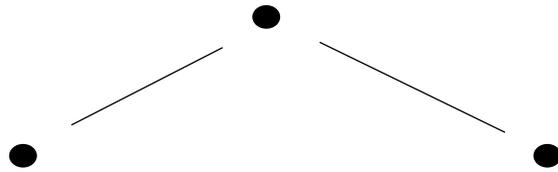
Was kann ich tun?

- Internationale Verflechtungen stärken (wirtschaftlich, wissenschaftlich, ...)
- Verbesserung Völkerverständigung, Austausch: kulturell, ...
- Vertrauen aufbauen, dazu nötig: Begegnungen, viele Kontakte
- Politische Kontakte
- Kontakte zu Medien
- ...

Was kann ich tun?

Andere überzeugen, exponentielles Wachstum nutzen: 2 weitere Personen:

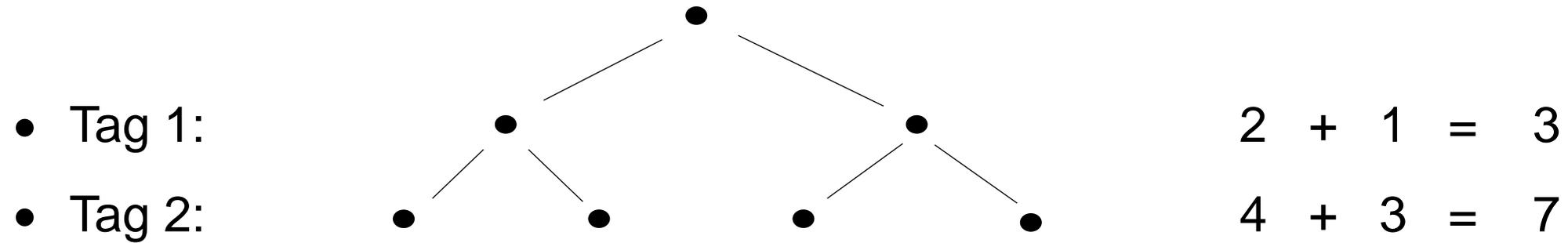
- Tag 1:



$$2 + 1 = 3$$

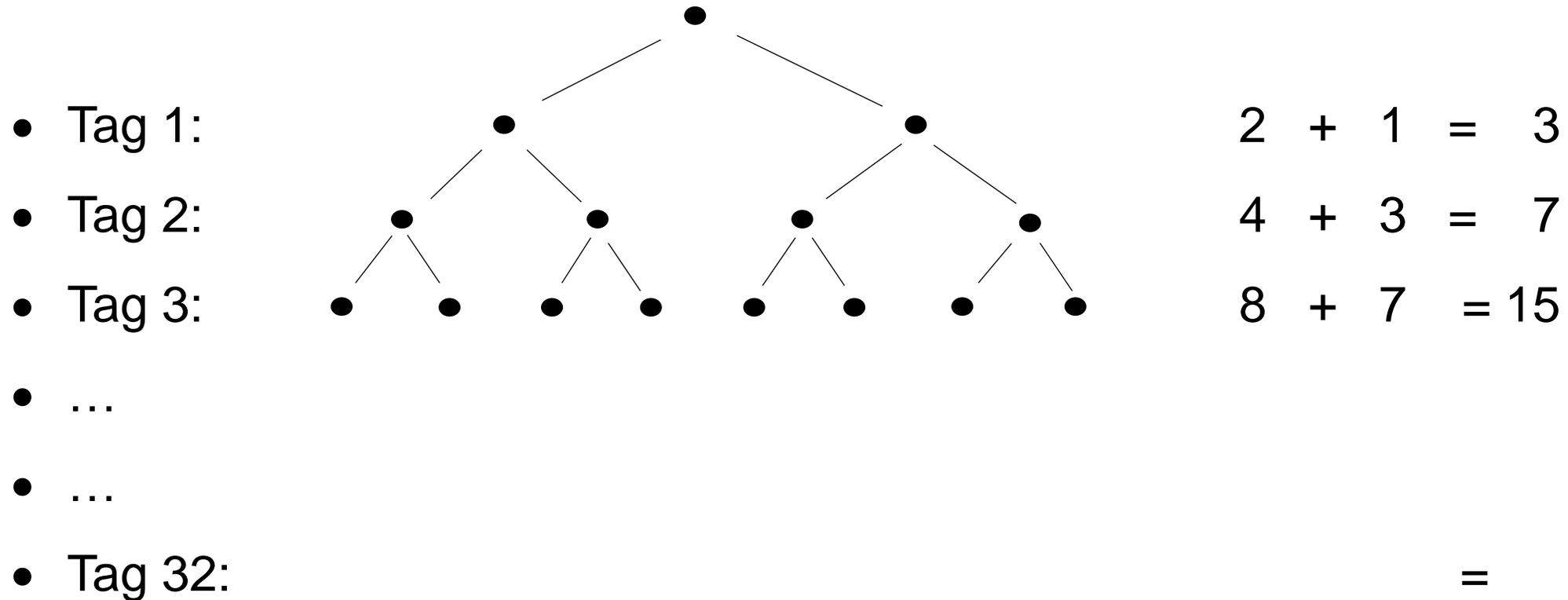
Was kann ich tun?

Andere überzeugen, exponentielles Wachstum nutzen: 2 weitere Personen:



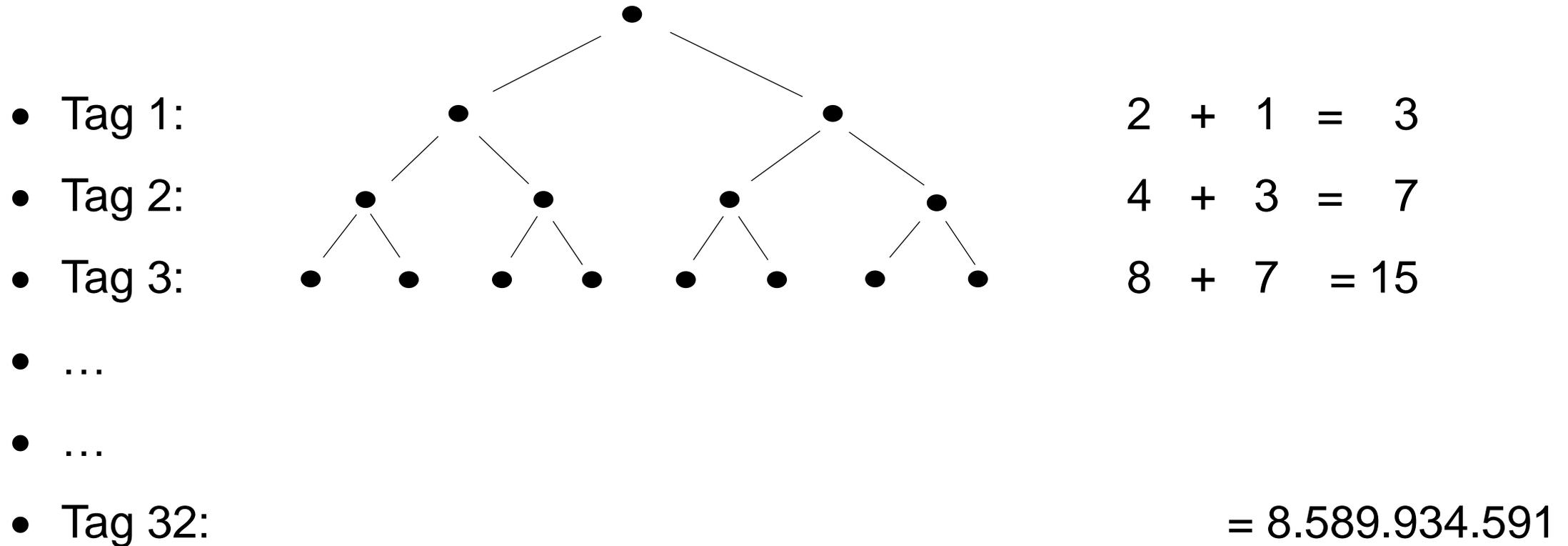
Was kann ich tun?

Andere überzeugen, exponentielles Wachstum nutzen: 2 weitere Personen:



Was kann ich tun?

Andere überzeugen, exponentielles Wachstum nutzen: 2 weitere Personen:



vielen Dank



akav.de



ki-folgen.de

Verwendete Fotos von Pixabay