Generative Künstliche Intelligenz und Cybersicherheit?

t

Marcus Braun

SentinelOne
Secure Tomorrow

Cybersecurity Summit

Generative KI?

Erste Instanz 1966!

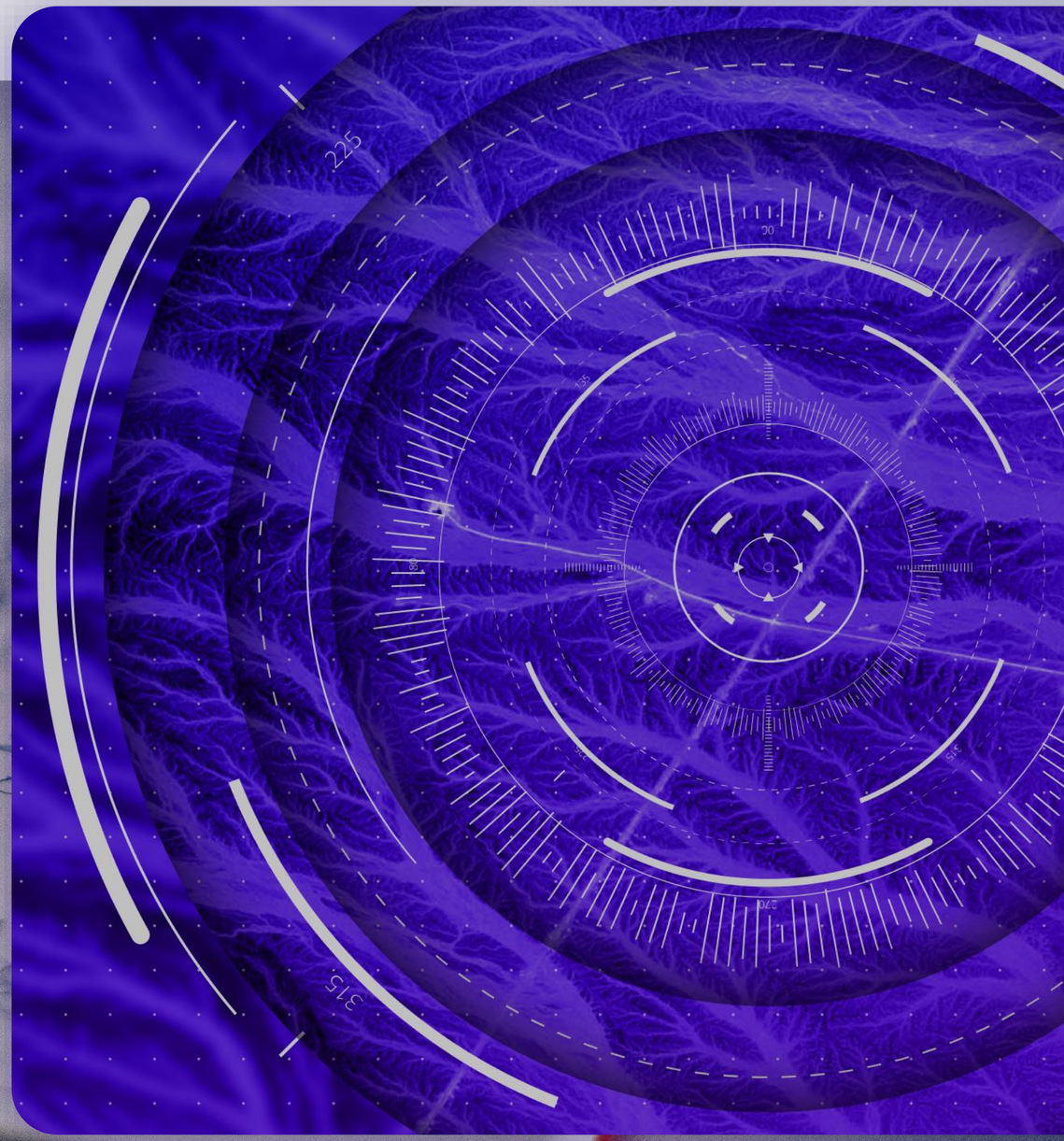
Chatbot ”ELIZA”, entwickelt am MIT. Sollte Konversationen mit einem Psychotherapeuten simulieren.

Automatisierte Prozesse

Dabei werden Algorithmen verwendet, um Daten (häufig in Form von Bildern oder Text) zu erstellen, zu manipulieren oder zu synthetisieren.

Erstellt etwas “Neues”

Erschafft auf Basis historischer Daten etwas, das es so noch nie gegeben hat.



Diskriminative vs. Generative KI

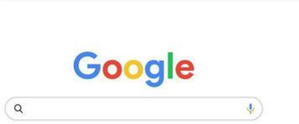
Diskriminative KI

möchte die Frage beantworten „Ist das ein Bild von einer Katze und einem Hund?“

Generative KI

reagiert auf eine Eingabe wie „Erstelle mir ein Bild von einem Hund mit einer Katze“.

ChatGPT



Indexbasierte Suche

Google hat die Art und Weise geprägt, wie moderne Internetnutzer das „Web“ entdeckten, konsumierten und sich darin bewegten.

- Aber es ähnelt dem Index eines Buches!
- Größtenteils manuell!
- Es liegt an dem Benutzer, die Treffer zu sichten!
- Einseitiges Gespräch!

ChatGPT

Indexbasierte Suche

Google hat die Art und Weise geprägt, wie moderne Internetnutzer das „Web“ entdeckten, konsumierten und sich darin bewegten.

- Aber es ähnelt dem Index eines Buches!
- Größtenteils manuell!
- Es liegt an dem Benutzer, die Treffer zu sichten!
- Einseitiges Gespräch!

„Sprechen“ in einer “natürlich klingenden” Sprache

Wenn eine Frage gestellt wird, greift ChatGPT auf eine riesige Menge an Textdaten zurück, darunter Bücher, Artikel und Webseiten, um seine Antwort zu generieren.

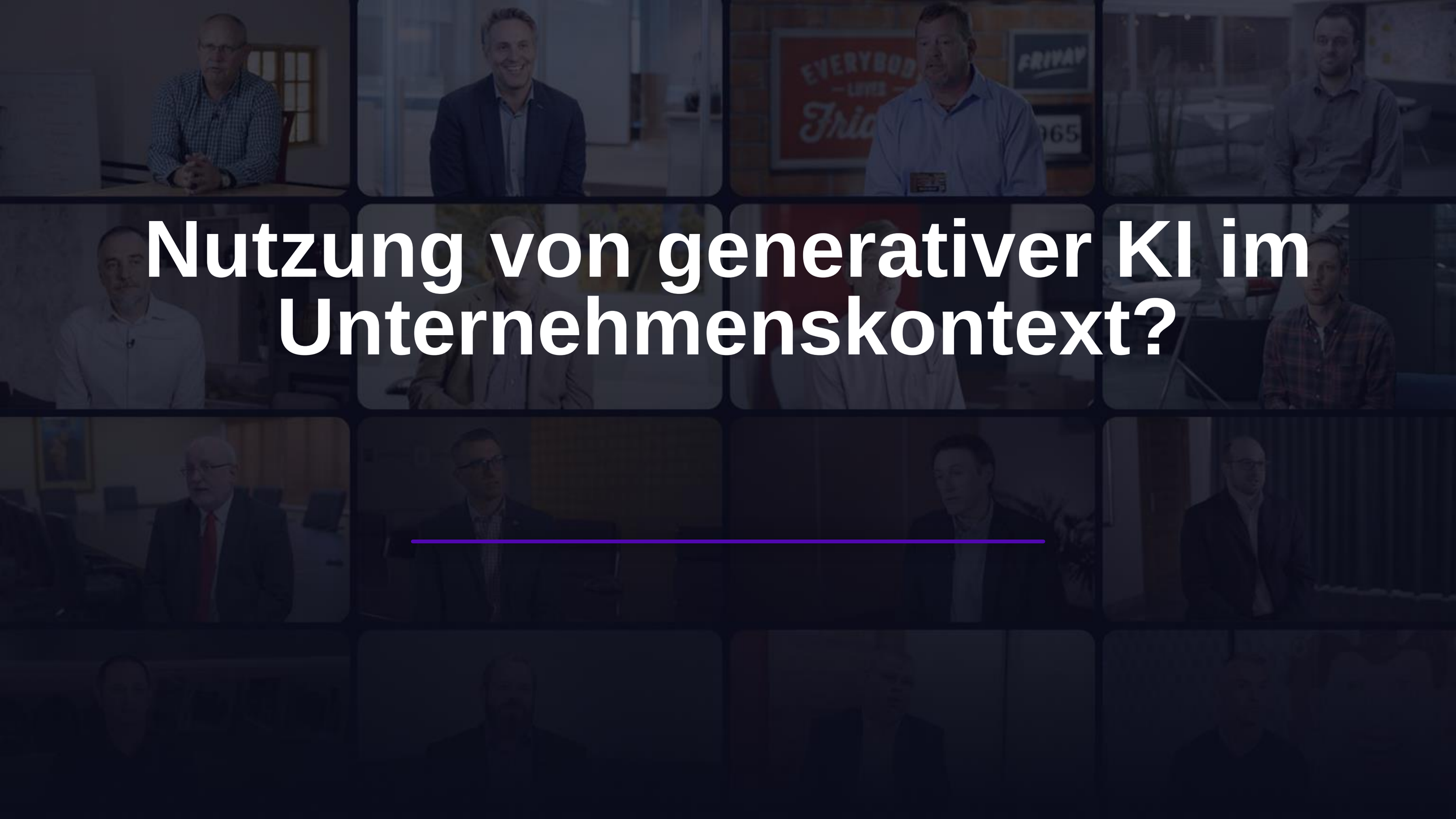
- Geschult, gesprächig zu sein!
- Verwendet maschinelle Lernalgorithmen, um im gleichen Kontext wie der Rest des Dialogs zu antworten!
- Antworten sind menschenähnlich!
- Geben dem Benutzer das vertraute Gefühl einer wechselseitigen Konversation!

Time Magazine

„ Der KI-Boom begann um das Jahr 2020 herum richtig Fahrt aufzunehmen, angetrieben durch mehrere entscheidende Durchbrüche im

- Design neuronaler Netzwerke, Large Language Models (LLMs)
- wachsende Datenverfügbarkeit,
- Bereitschaft von Technologieunternehmen, für gigantische Rechenleistung zu zahlen.“

https://time.com/6255952/ai-impact-chatgpt-microsoft-google/



Nutzung von generativer KI im Unternehmenskontext?

Nützliches Tool oder böse Waffe?

Risiko oder Chance?

Generelle Herausforderungen mit KI

Gültig und zuverlässig; rechenschafts-pflichtig und transparent

Vertrauenswürdigkeit des Systems

Regulatorische Unsicherheiten

Balance Risiken und Akzeptanz

Ressourcen

Compute und Expertise

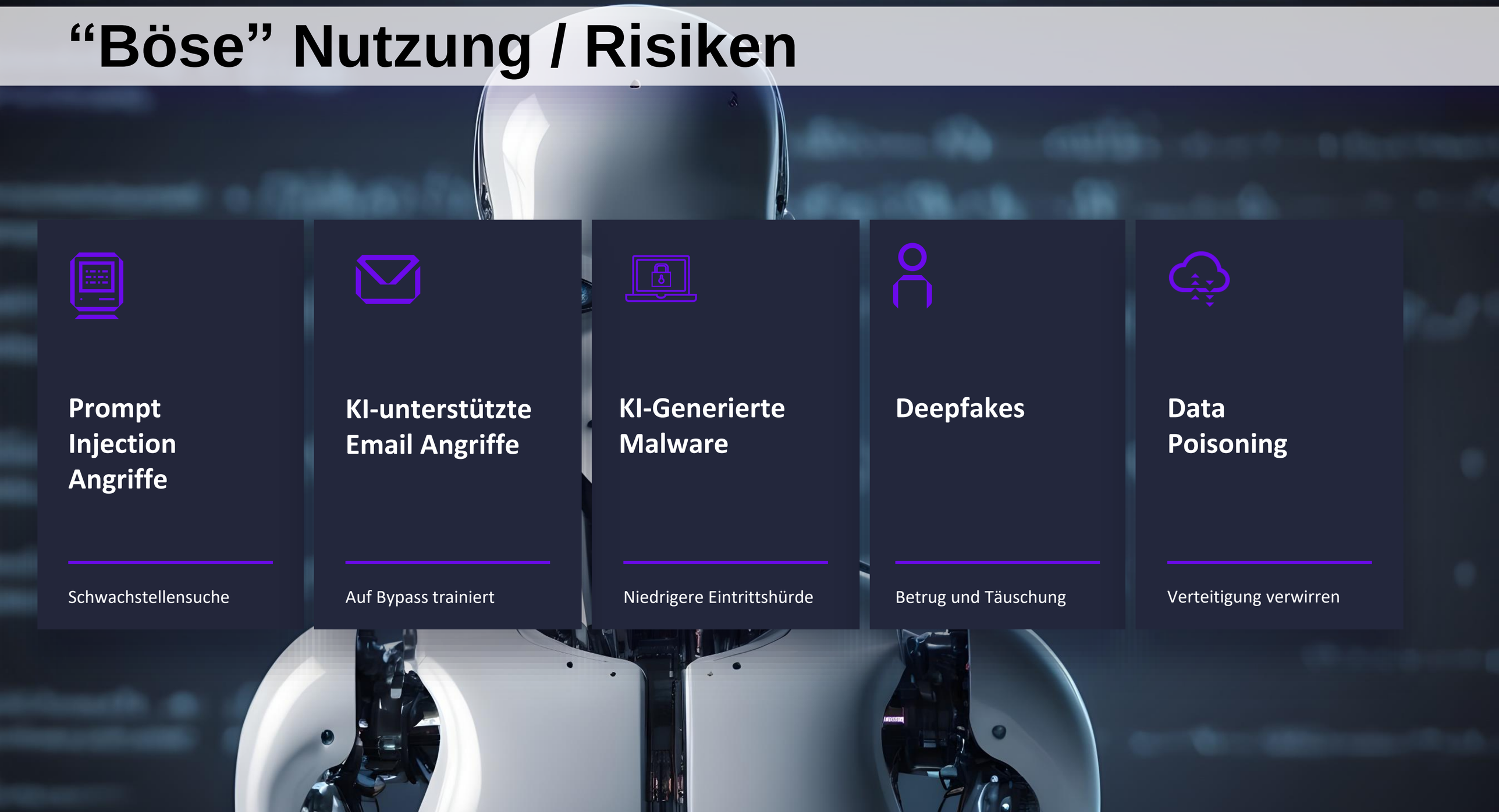
Genauigkeit

Klar definiert und realistisch

Sicherheit, Privatsphäre, Voreinge-nommenheit, ethisch?

Eigenschaften

“Böse” Nutzung / Risiken



Prompt Injection Angriffe

Schwachstellensuche

KI-unterstützte Email Angriffe

Auf Bypass trainiert

KI-Generierte Malware

Niedrigere Eintrittshürde

Deepfakes

Betrug und Täuschung

Data Poisoning

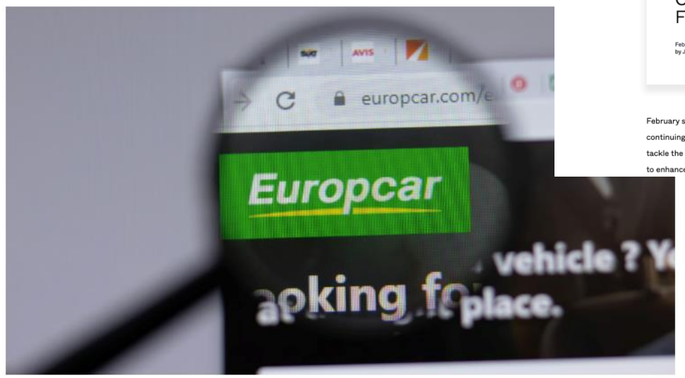
Verteitigung verwirren

“Böse” Nutzung / Risiken

Europcar dismisses data leak claims as AI fake

Updated on: January 31, 2024 10:49 AM

Vilius Petkauskas, Deputy Editor



SentinelOne

February 2024 Cybercrime Update | Commercial Spyware, AI-Driven APTs & Flawed RMMs

February 27, 2024
by Jim Walter

February saw the U.S. government take significant actions against cybercrime, continuing the current administration's policy of using all the resources of the state to tackle the problem head on. Nation-state actors, meanwhile, have taken to leveraging AI to enhance their operations and attacks.

SECURITY

Football Australia details

by Vilius Petkauskas

Australia's football gc keys potentially open including players' do

CNET

Tech Money Home Wellness Home Internet

News > Misinformation

Misinformation: How It Works d Ways to Spot It

rmining what's real online is getting more difficult as AI deepfakes spread across social media platforms. But e are steps you can take to deal with it.

Listen to this Post

Table of Contents

Oscar Gonzalez

Nov. 8, 2023 2:01 p.m. PT

SC Media

CISO STORIES TOPICS EVENTS PODCASTS RESEARCH RECOGNITION LEADERS

ADVERTISEMENT

RSAConference2024 San Francisco | May 6 – 9 | Moscone Center

CYBERSECURITY'S BIGGEST EVENT AWAITS YOU.

REGISTER NOW

REGISTER BY APR. 5 AND SAVE UP TO $60

AI benefits/risks, Generative AI

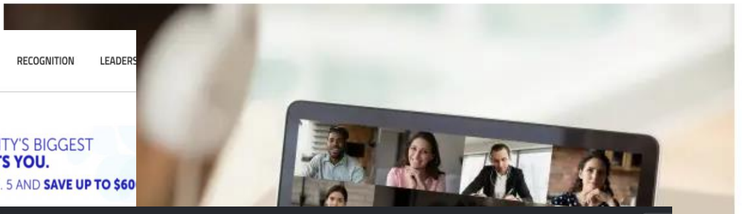
Google launches AI bug bounty pr organizations plan to study risks

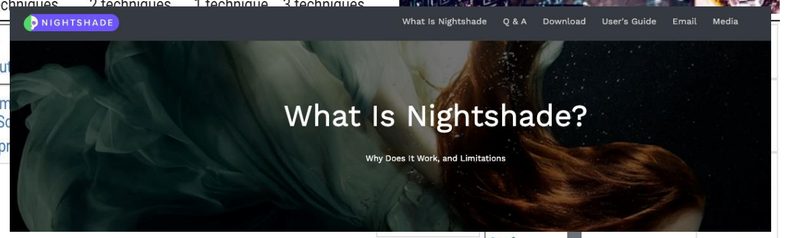
Steve Zurier October 27, 2023

Videokonferenz voller KI-Klone: Angestellter schickt Betrügern 24 Millionen Euro

Bislang werden im Rahmen der "Chef-Masche" Angestellte zumeist von einer Person überzeugt, Geld herauszugeben. Ein Fall in Hongkong hat nun eine neue Qualität.

Lesezeit: 2 Min. In Pocket speichern 65





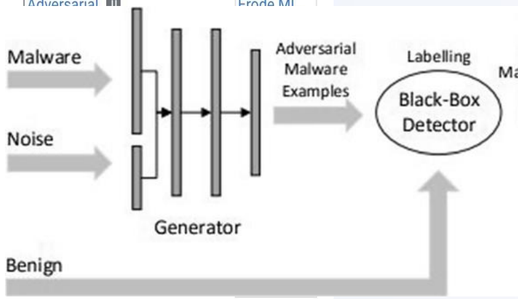
Reconnaissance &	Resource Development &	Initial Access &	ML Model Access	Execution &	Persistence &	Defense Evasion &	Discovery &
5 techniques	7 techniques	4 techniques	4 techniques	2 techniques	2 techniques	1 technique	3 techniques
Search for Victim's Publicly Available Research Materials	Acquire Public ML Artifacts	ML Supply Chain Compromise	ML Model Inference API Access	User Execut			
Search for Publicly Available Adversarial Vulnerability Analysis	Obtain Capabilities &	Valid Accounts &	ML-Enabled Product or Service	Comm and Sc			
Search Victim-Owned Websites	Develop Adversarial ML Attack Capabilities	Evade ML Model	Physical Environment Access	Interpr			
Search Application Repositories	Acquire Infrastructure	Exploit Public-Facing Application &	Full ML Model				
Active Scanning &							

Craft Adversarial | Erode ML

BlackMamba ChatGPT Polymorphic Malware | A Case of Scareware or a Wake-up Call for Cyber Security?

March 16, 2023
by Migo Kedem

PDF



StavC / ComPromptMized Public

Code Issues 1 Pull requests Actions Projects Security Insights

master 1 Branch 0 Tags Go to file Code

StavC 1.Adding the ArXiv Paper Link. 686c3d6 · 2 months ago 3 Commits

.idea	1.Adding the ArXiv Paper Link.	2 months ago
Assets	inital commit	3 months ago
FlowSteering	Initial commit	3 months ago
RAG-based Worm	inital commit	3 months ago
README.md	1.Adding the ArXiv Paper Link.	2 months ago
requirements.txt	inital commit	3 months ago

README

ComPromptMized: Unleashing Zero-click Worms that Target GenAI-Powered Applications

NIST

NEWS

NIST Identifies Types of Cyberattacks That Manipulate Behavior of AI Systems

Publication lays out “adversarial machine learning” threats, describing mitigation strategies and their limitations.

January 04, 2024

- AI systems can malfunction when exposed to untrustworthy data, and attackers are exploiting this issue.
- New guidance documents the types of these attacks, along with mitigation approaches.
- No foolproof method exists as yet for protecting AI from misdirection, and AI developers and users should be wary of any who claim otherwise.

MEDIA CONTACT
Chad Boutin
charles.boutin@nist.gov
(301) 975-4261

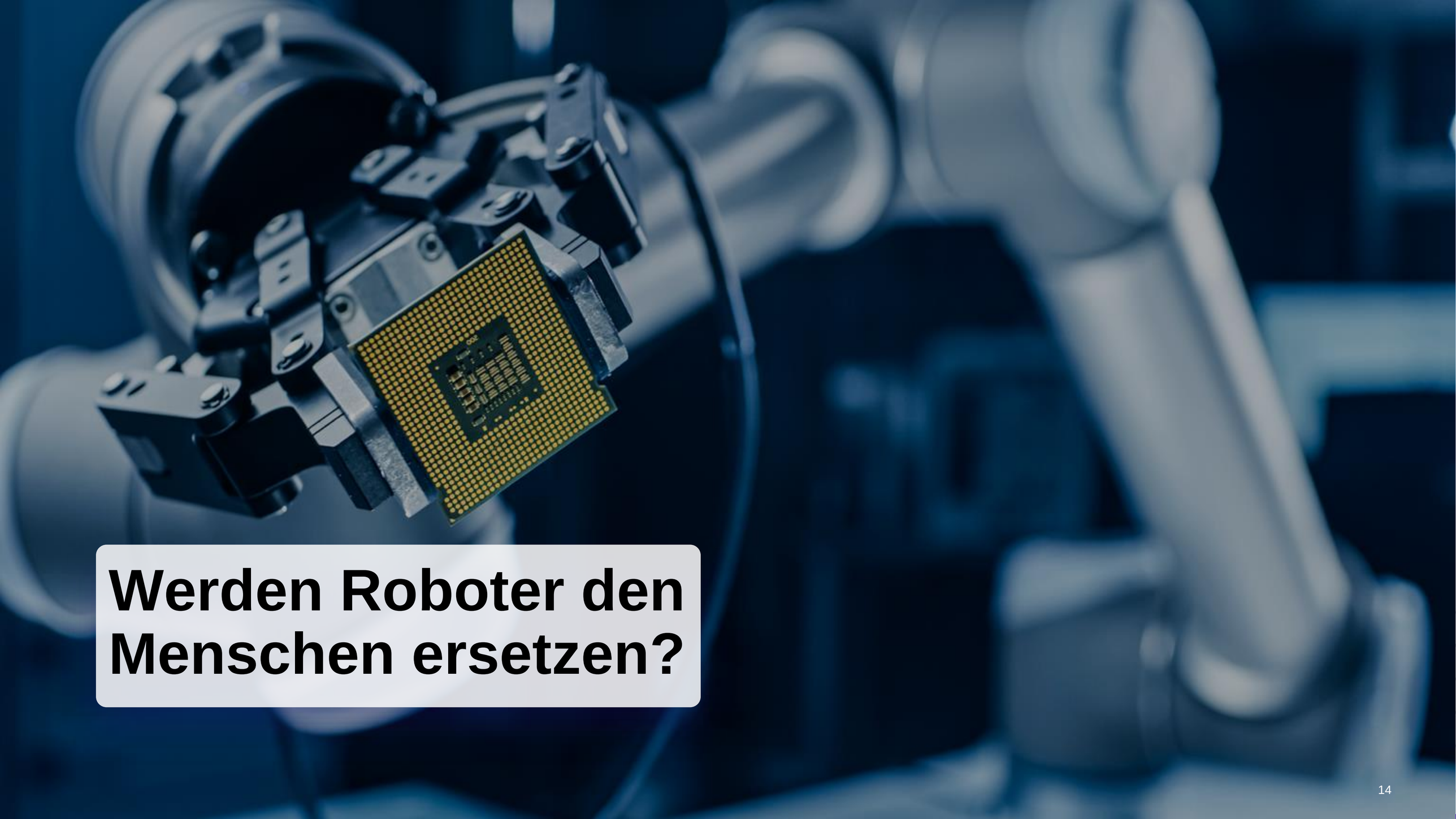
ANALY

ver,
ss.
nas
their

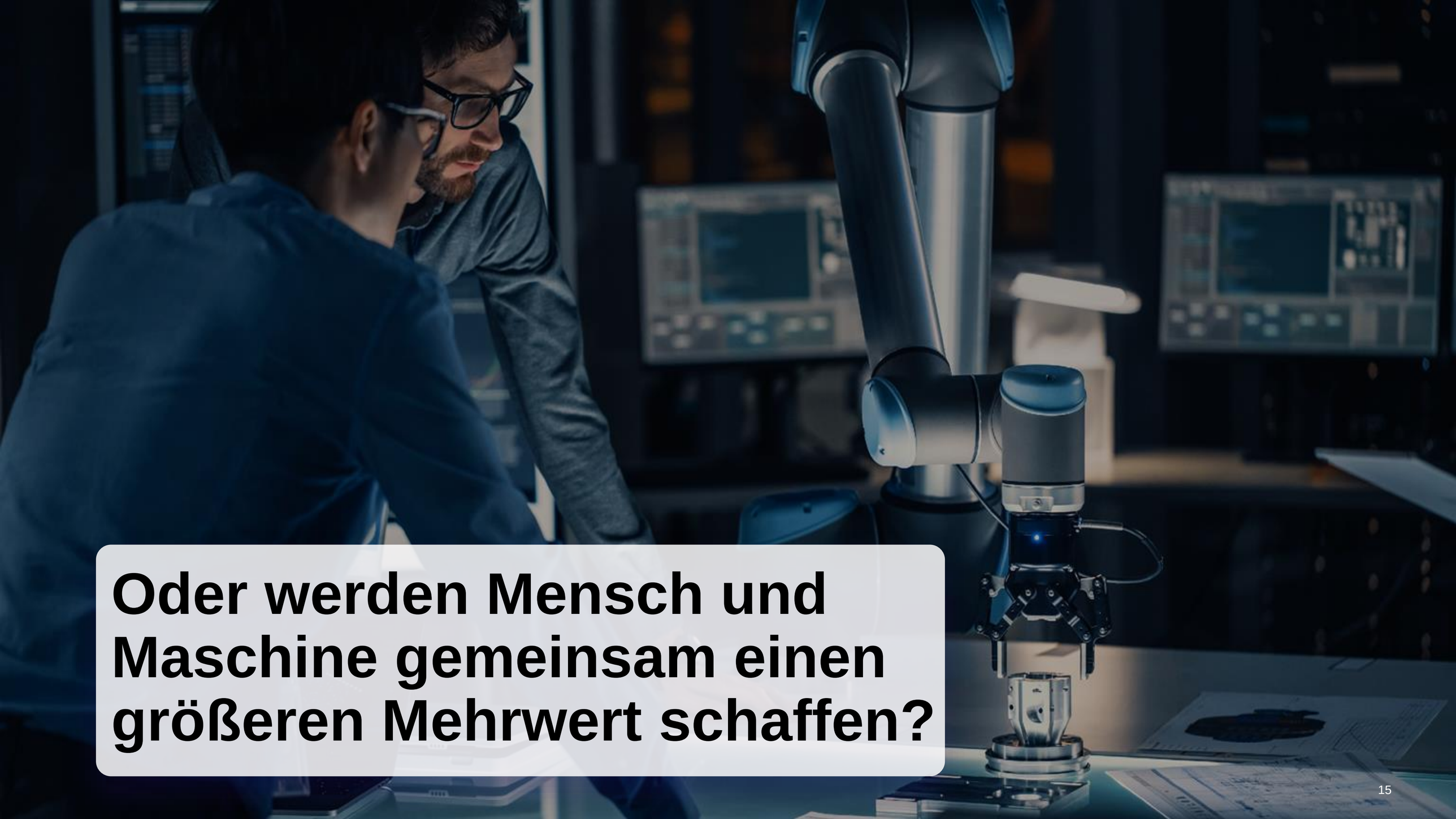
ible, as the race for innovation pushes
ion of AI-driven productivity tools and
es and vectors.

Je nach Nutzung

Offen sein
Verstehen
Bewerten



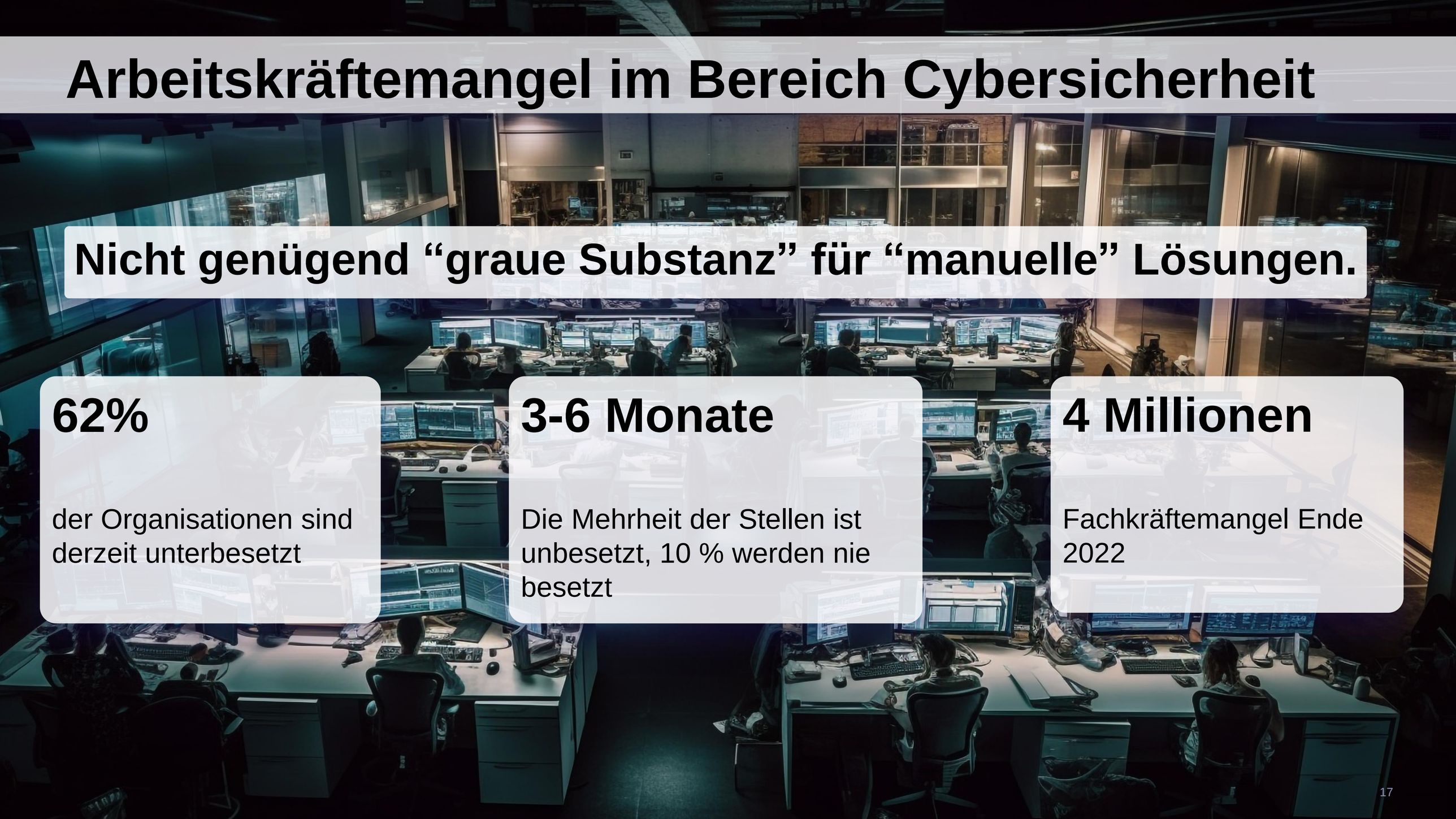
Werden Roboter den Menschen ersetzen?



Oder werden Mensch und Maschine gemeinsam einen größeren Mehrwert schaffen?

Unterstützung von Cybersicherheitsanalysten durch KI-gesteuerte Bedrohungssuche, -analyse und –reaktion?

Arbeitskräftemangel im Bereich Cybersicherheit



Nicht genügend “graue Substanz” für “manuelle” Lösungen.

62%

der Organisationen sind derzeit unterbesetzt

3-6 Monate

Die Mehrheit der Stellen ist unbesetzt, 10 % werden nie besetzt

4 Millionen

Fachkräftemangel Ende 2022

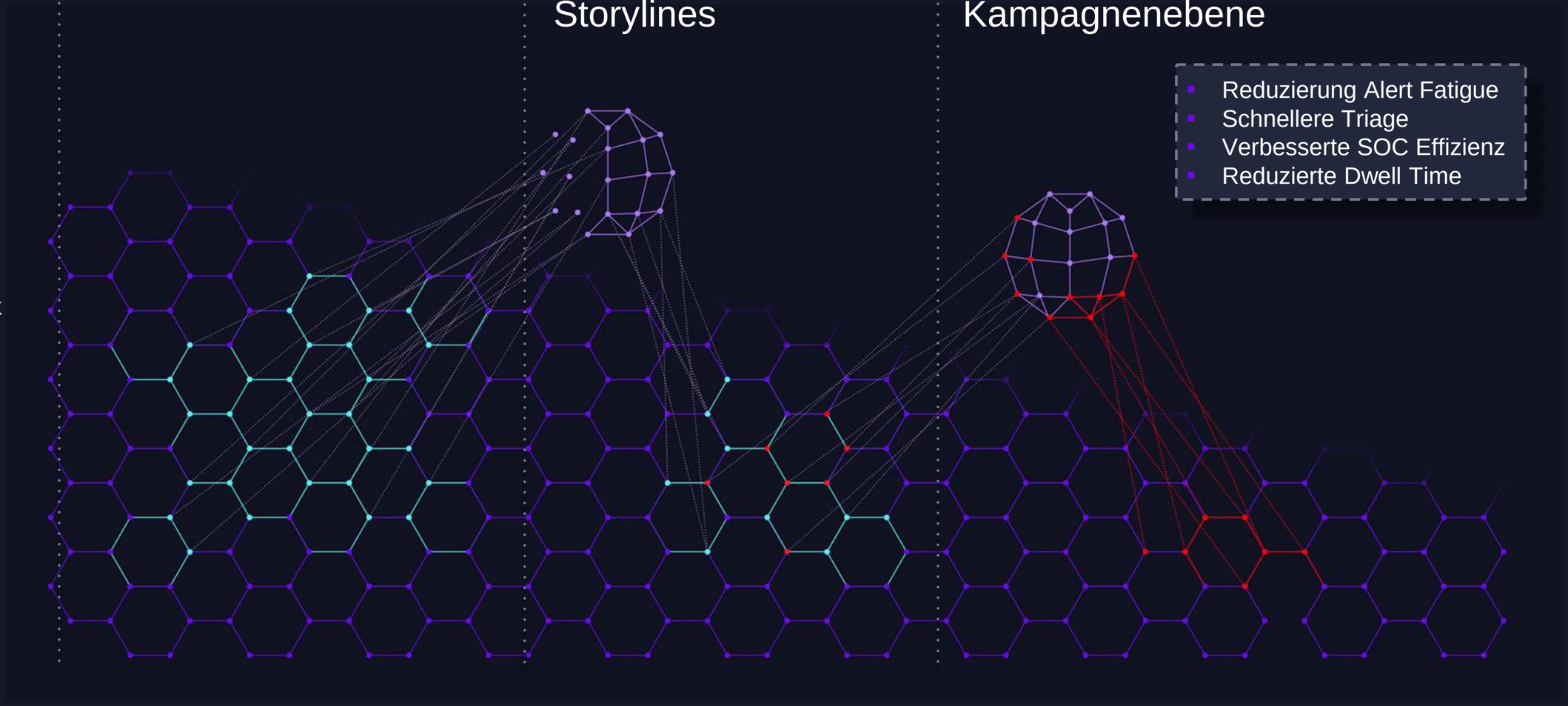
Signal : Noise Reduction ist kritisch

Trillions
Zeilen mit Rohdaten

Millionen von
angereicherten Storylines

Handvoll umsetzbarer
Vorfälle auf Kampagnenebene

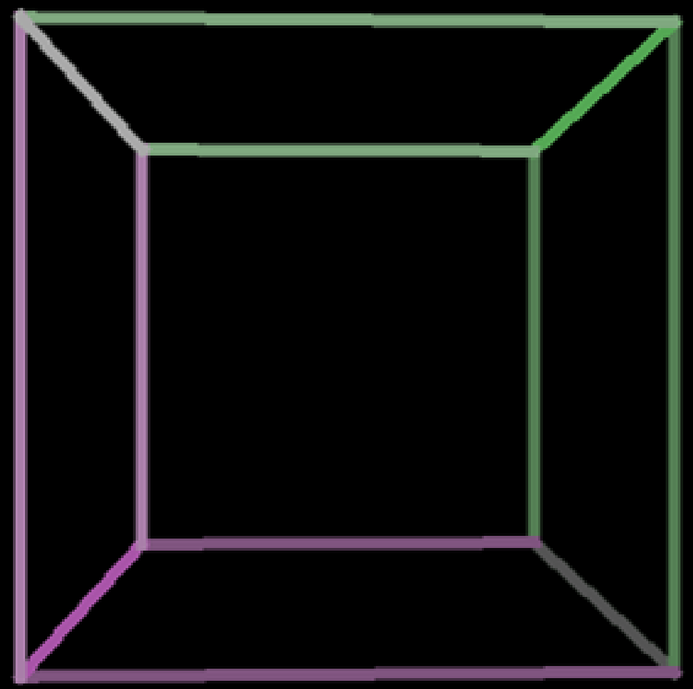
- Endpoint
- Email
- SIEM
- Netzwerk
- Firewall
- Cloud
- Identity



- Reduzierung Alert Fatigue
- Schnellere Triage
- Verbesserte SOC Effizienz
- Reduzierte Dwell Time

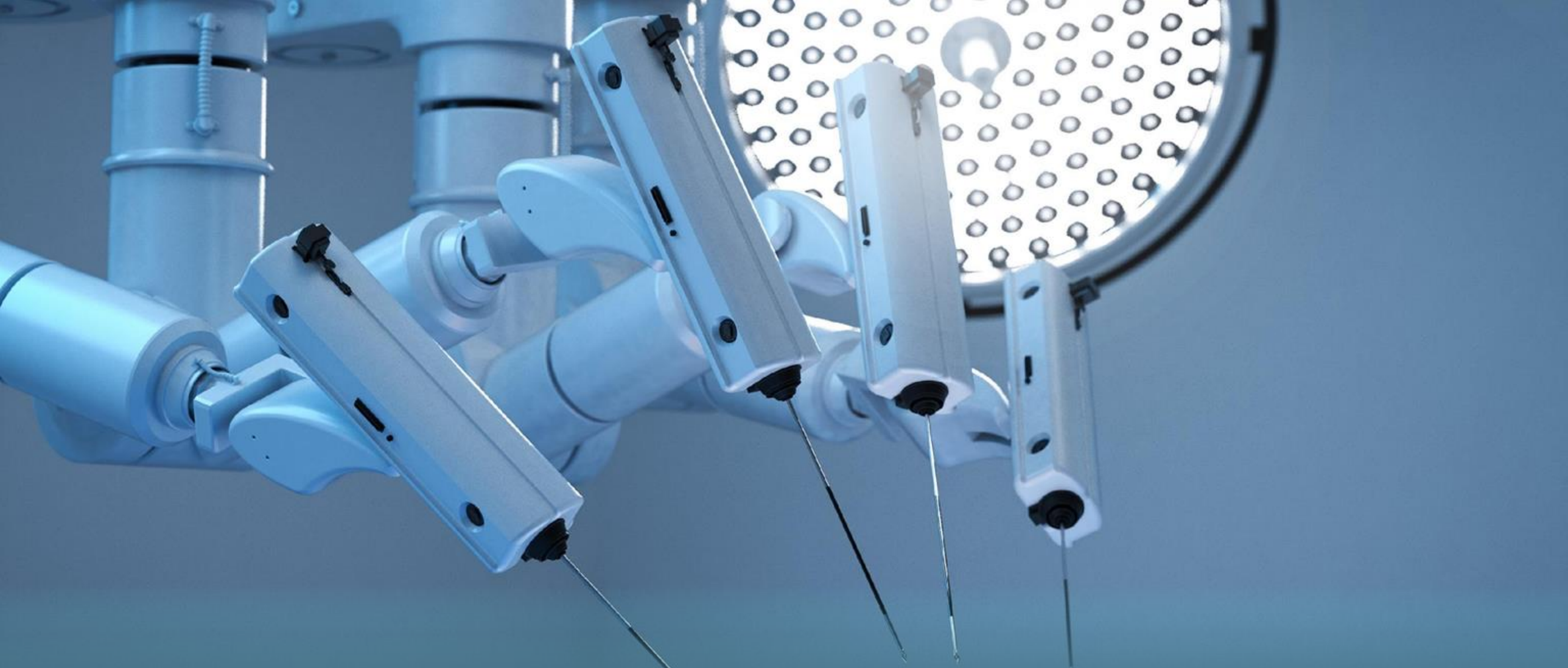
„Dualität“ der Black-Box-KI

- Nicht deterministisch (Ergebnisse können variieren)
- Erkennt bisher unerkannte Angriffe
- Oft nicht möglich, „Begründung“ für die Schlussfolgerung zu erklären
- False Positive und False Negative sind inhärent



Genauigkeit und Erklärbarkeit?

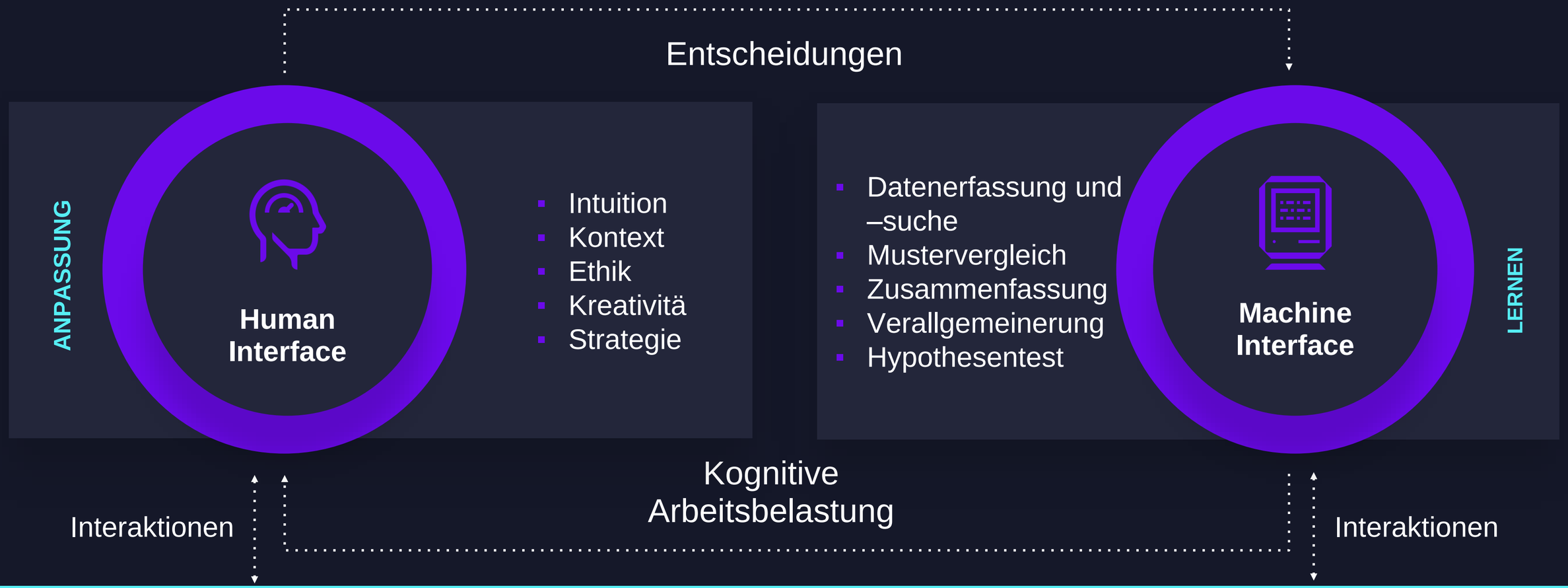
Kann KI den Unterschied erkennen?



Fühlen Sie sich in diesem Szenario mit einer Genauigkeit von 90 % genauso wohl?

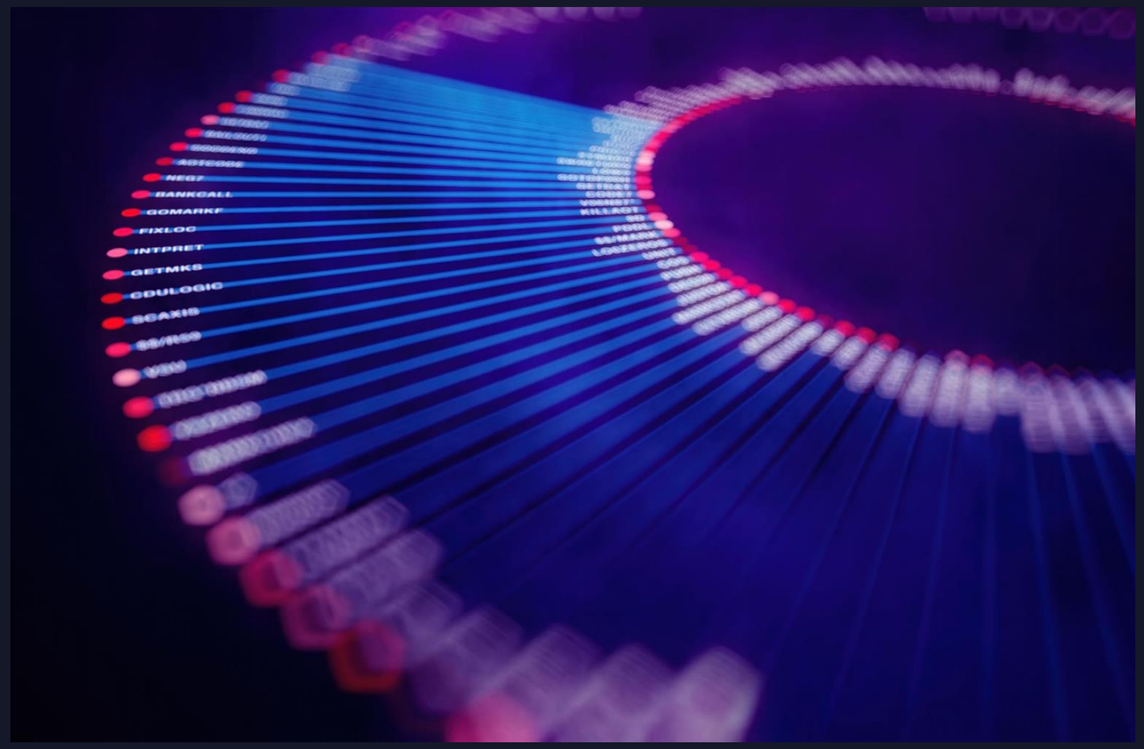
Fühlen Sie sich in diesem Szenario mit einer Genauigkeit von 90 % genauso wohl?

Eine Zeit und ein Ort für „Maschinen“



Das Ganze ist mehr als die Summe seiner Teile

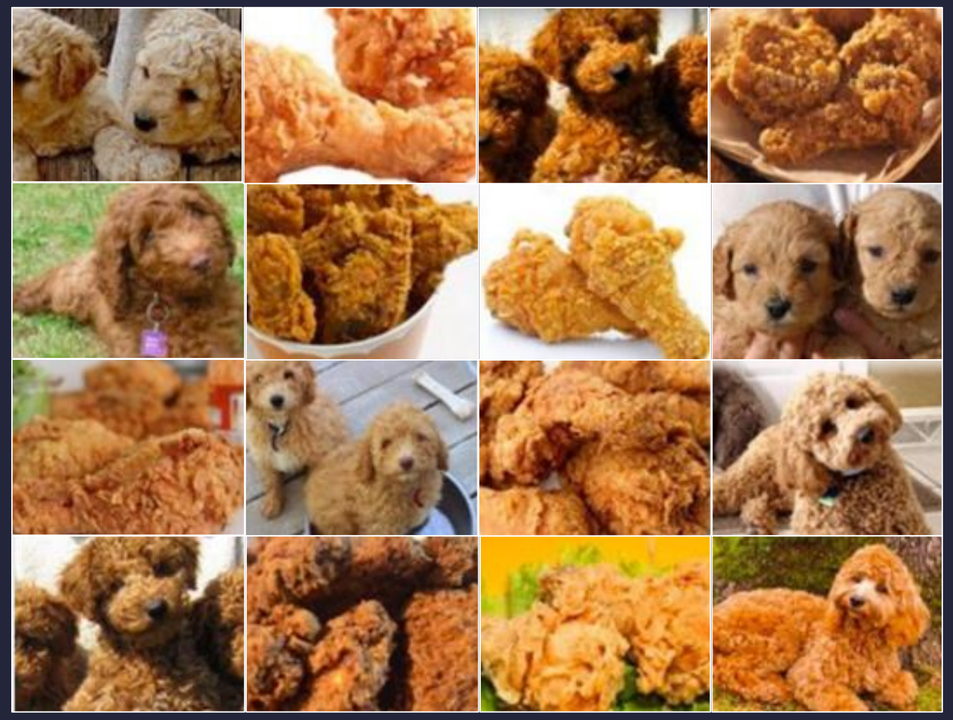
Komplexe Daten Analyse



Menschliches intellektuelles Kapital

Das Ganze ist mehr als die Summe seiner Teile

AI: 90% Akkurat



Fried Chicken oder Labradoodle?

Menschen: 100% Akkurat

Chocolate Chip Ice Cream oder Dalmatiner?

Mensch und Maschine sind **Besser Zusammen**!

Die zentralen Thesen

Immenses Potential

Potential

Verbessert die Cybersicherheit erheblich

Verbesserung

Menschen + AI

Ganzheitlich

Verbessern Sie die defensive KI kontinuierlich

Verteitigung

Zusammenarbeit ist für die Wirksamkeit von entscheidender Bedeutung

Fähigkeiten


Purple ai