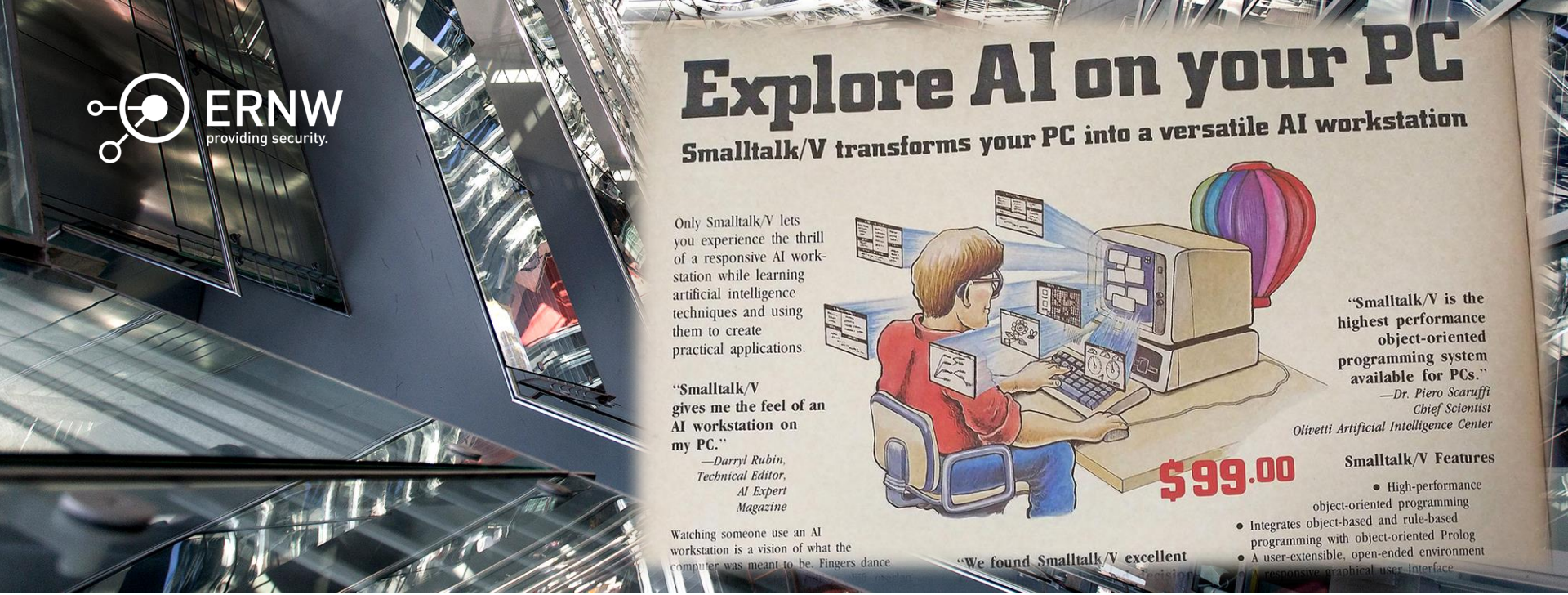


Computer Language Magazine, September 1986

IT Sicherheit und KI-Anwendungen – die neuen Herausforderungen meistern

Florian Grunow fgrunow@ernw.de
Hannes Mohr hmohr@ernw.de

ERNW
providing security.

Agenda

- AI und Angriffe
- AI und Verteidigung
- Sicherheit von AI-Anwendungen
- AI im Unternehmen
- Fragen und Diskussion

ERNW
providing security.

Suumit Shah
@suumitshah

We had to layoff 90% of our support team because of this AI chatbot.

Tough? Yes. Necessary? Absolutely.

The results?

Time to first response went from 1m 44s to INSTANT!
Resolution time went from 2h 13m to 3m 12s
Customer support costs reduced by ~85%

Here's how's we did it 🧵

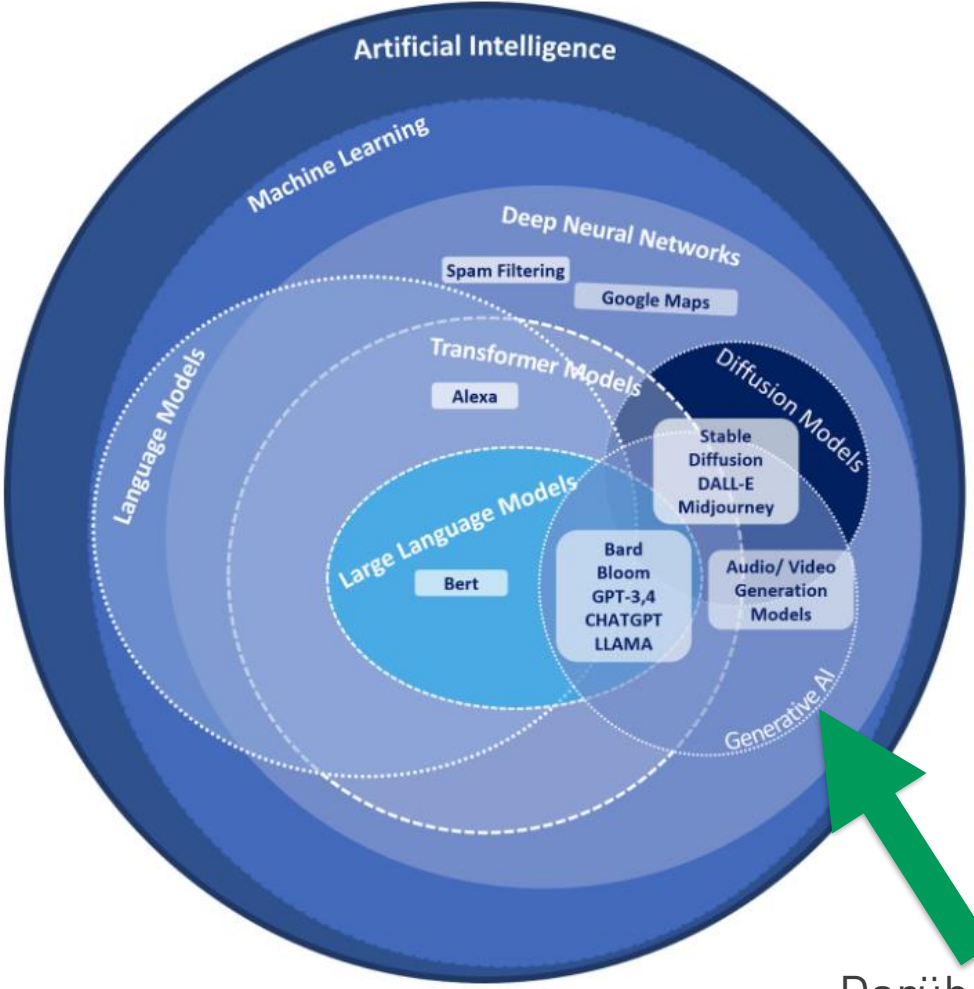
1:45 PM · Jul 10, 2023 · **2M** Views

MIT News
https://news.mit.edu › npr-227 · Diese Seite übersetzen

Clip NPR 10 reasons why AI may be overrated (Aug 6)

06.08.2024 — Acemoglu notes he believes **AI is overrated because humans are underrated**. "A lot of people in the industry don't recognize how versatile, talented, ...

Technologien



Darüber reden alle

Figure 1.1: Image of LLM relationship within the field of Artificial Intelligence
https://owasp.org - LLM AI Security & Governance Checklist

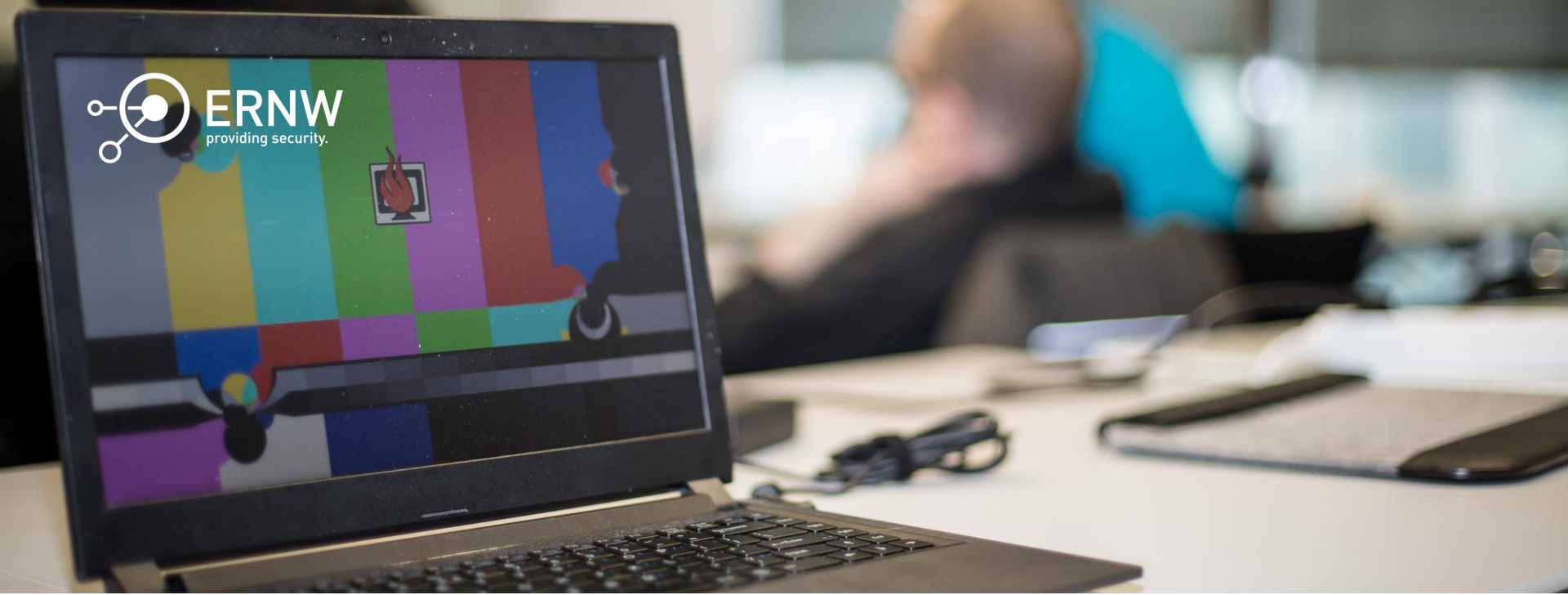
AI und Angriffe

ERNW
providing security.

AI und Angriffe

Im Wesentlichen lassen sich hier zwei Punkte unterscheiden

- AI unterstützt Angriffe (Gegen Maschinen)
- AI ist der Angriff (Gegen Menschen)



AI unterstützt Angriffe

Beispiele...

ERNW
providing security.

AI unterstützt Angriffe 1/2

Ausgangssituation:

- Produkt hat Sicherheitslücke
- Hersteller veröffentlicht Patch

Hergang:

- Angreifer analysiert (mit AI) Patch
- Angreifer entwickelt (mit AI) Schadcode
- Angreifer nutzt Sicherheitslücke aus

Bildquelle:
https://www.pda.org/pda-letter-portal/home/full-article/add-sherlock-holmes-to-your-investigation-team

AI unterstützt Angriffe 2/2

Wann kann das klappen?

“Nur”, wenn man zu langsam Patched

Wie schnell waren Angreifer denn ohne AI?

It depends, 1 bis N-Tage

Bildquelle:
https://www.pda.org/pda-letter-portal/home/full-article/add-sherlock-holmes-to-your-investigation-team

ERNW
providing security.

Ist das also gar nicht interessant?

Doch ☺

Wenn der Angreifer stattdessen folgende Frage stellt:

Welche Patches, die nicht als security relevant markiert sind, sind in Wahrheit doch security relevant?

Dann wird es interessanter!

Bildquelle:
https://www.freepik.com/free-vector/question-mark-exclamation-mark-brush-stroke_49610162.htm

ERNW
providing security.

Sind wir jetzt verloren?

Nein ☺

Das war schon immer ein Problem…

- Patchen möglichst schnell
- Wenn möglich auch Patchen, wenn es nicht als Security-relevant markiert ist
- Bzgl. Der Bedrohung durch unbekannte Schwachstellen: Mehrere Verteidigungslinien umsetzen

Bildquelle:
https://www.freepik.com/free-vector/question-mark-exclamation-mark-brush-stroke_49610162.htm

ERNW
providing security.

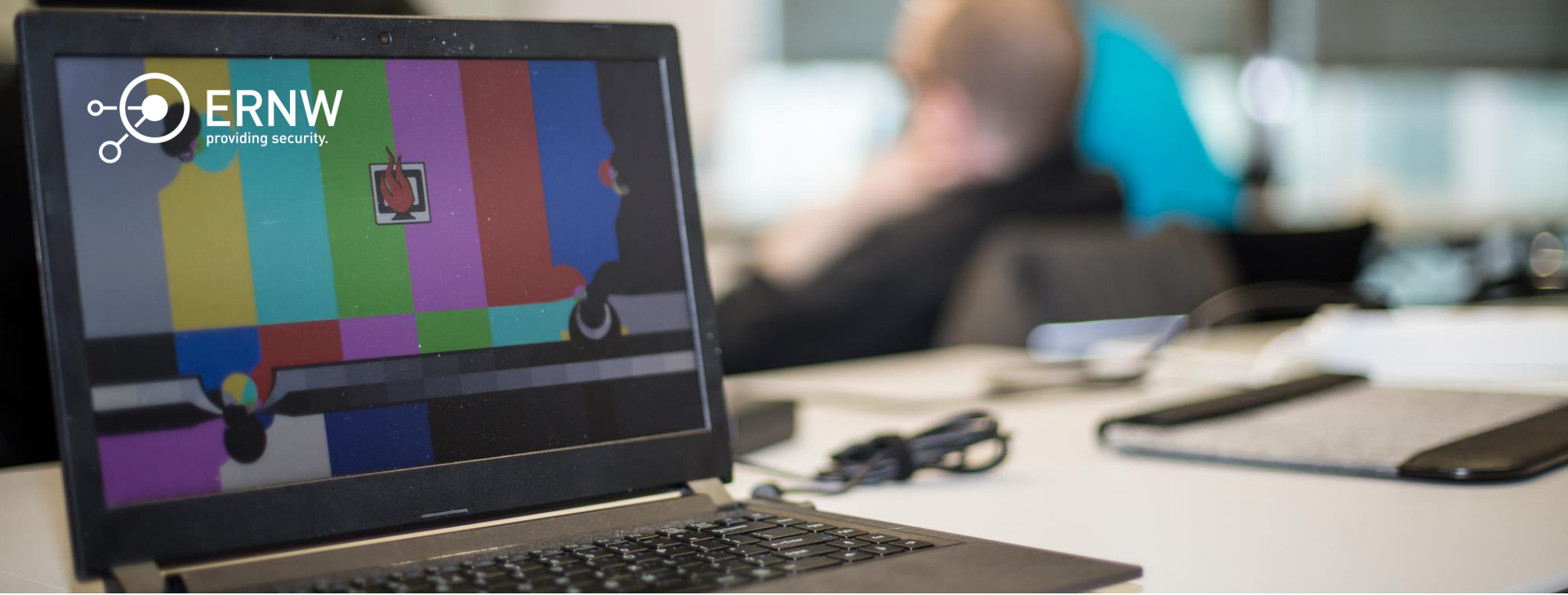
Fazit – AI unterstützt Angriff

Alles nur halb so schlimm…
… aber schon auch schlimm

Das ist alles nichts fundamental Neues!

Die gute Nachricht:
Man muss nichts Neues machen (wegen AI)

Die schlechte Nachricht:
Man sollte etwas machen (nicht nur wegen AI)



AI ist der Angriff

Beispiele...

AI ist der Angriff 1/2

Ausgangssituation:

Unternehmen beschäftigen Menschen

Problem:

Menschen sind keine Maschinen (und machen Fehler)

Hergang:

Angreifer verwendet Social Engineering(-AI)

Menschen lassen sich (durch die AI) zu Aktionen überreden

Bildquelle:
https://www.pda.org/pda-letter-portal/home/full-article/add-sherlock-holmes-to-your-investigation-team

AI ist der Angriff 1/2

Wann kann das klappen?

“Nur”, wenn man Menschen anstellt ;)

Ging das auch ohne AI?

Ja

Bildquelle:
https://www.pda.org/pda-letter-portal/home/full-article/add-sherlock-holmes-to-your-investigation-team

ERNW
providing security.

Und nun?

Menschen sind keine Maschinen …
… sie können Vorgänge hinterfragen

Im Gegensatz zu AI (bis jetzt)!

Bildquelle:
https://www.freepik.com/free-vector/question-mark-exclamation-mark-brush-stroke_49610162.htm

ERNW
providing security.

Fazit – AI ist der Angriff

Alles nur halb so schlimm…
… aber schon auch schlimm

Man muss eigentlich nichts neues machen (wegen AI)

Man sollte zum Beispiel Prozesse haben die verhindern, dass per Telefon Überweisungen veranlasst werden

Bildquelle:
https://www.freepik.com/free-vector/question-mark-exclamation-mark-brush-stroke_49610162.htm

AI und Verteidigung

ERNW
providing security.

Waren Sie z.Bsp. auf der IT-SA letztes Jahr?

AI machen jetzt alle – AI Powered everything

- Network Security…
- Cloud Security…
- Endpoint Security…
- Mobile Security…
- IoT Security…
- Application Security…
- Zero Trust…
- DNS…

… erinnert ein bisschen an Block Chain

ERNW
providing security.

Wofür eignet sich AI gut?

AI machen jetzt alle – AI Powered everything

- Verknüpfungen herstellen
- Große Datenmengen erfassen

ERNW
providing security.

Was für Probleme hat AI?

- Ergebnisse und Herleitung oft nur schwer nachvollziehbar
- Verlässlichkeit lässt sich nur schwer erfassen
- Es ist ein weiterer Layer Abstraktion
- Es ist ein weiterer angreifbarer Teil

ERNW
providing security.

Fazit – AI als Security Produkt

- Whitelists sind besser als Blacklists
- Explizit ist besser als implizit
- Alerts, Logs und Backups sind nur sinnvoll, wenn man sie auch benutzen kann
- Operations is key!
- Es ist besser Produktsicherheit herzustellen, als Sicherheit als Produkt zu kaufen

AI-Produkte und Sicherheit

ERNW
providing security.

Angriffsoberfläche von AI Anwendungen

- Sehr viel Verwirrung!
- Wichtige Unterscheidung
  - “application/infrastructure” security -> Nichts Neues!
  - “AI” security -> Was gibt es hier?

ERNW
providing security.

AI und Angriffe

Im Wesentlichen lassen sich hier zwei Punkte unterscheiden

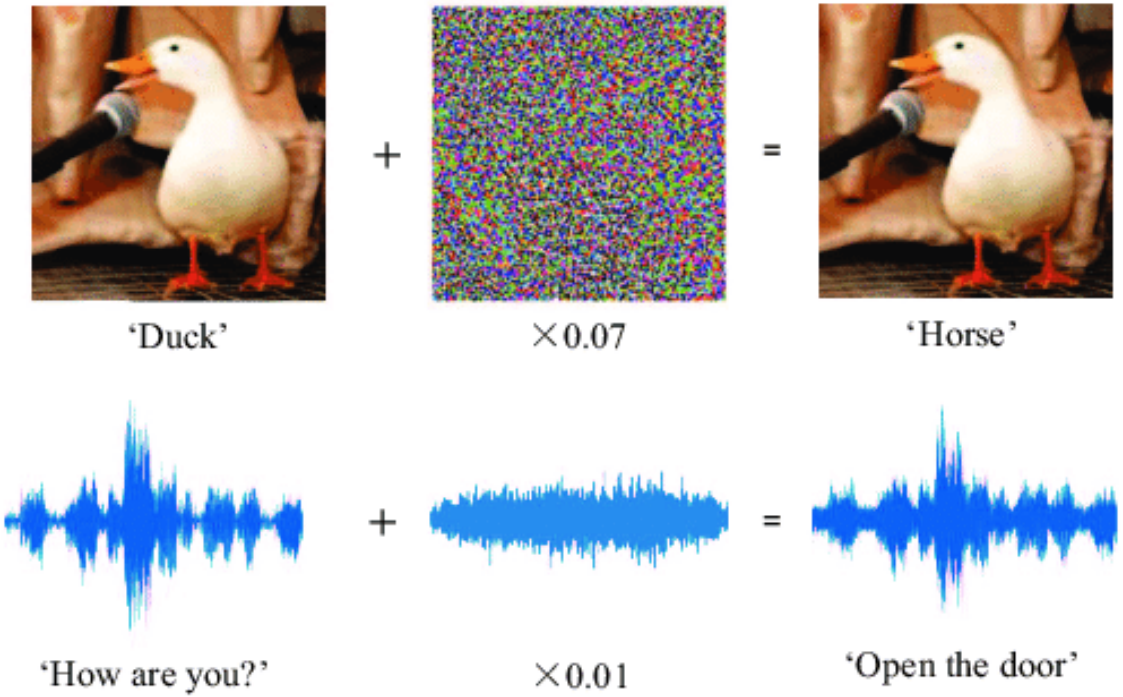
- AI Angriffe (klassisch)
- AI Angriffe

“As soon as it works, no one calls it AI anymore”

- John McCarthy

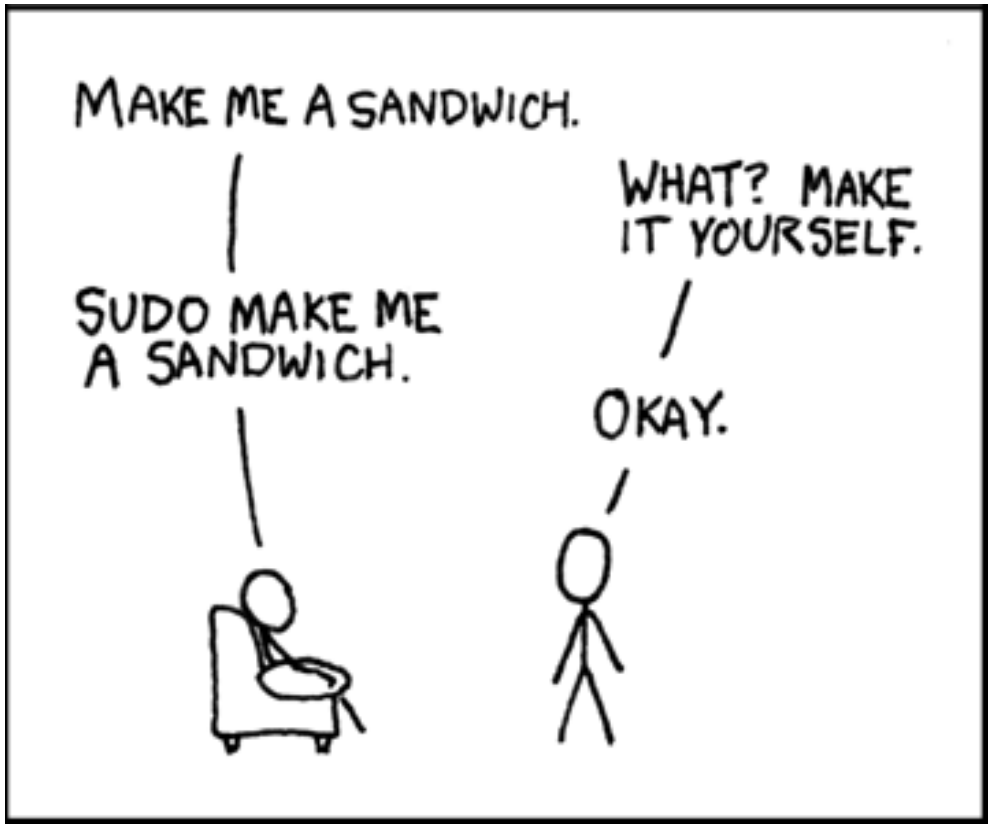
AI Angriffe – Gegen klassische AI

Zum Beispiel Adversarial Samples



Quelle:
https://collab.dvb.bayern/spaces/TUMdlma/pages/73379903/Neural+Network+Robustness+adversarial+examples

AI Angriffe – Gegen LLMS



Zum Beispiel Prompt Injection

Quelle: https://xkcd.com/149/

ERNW
providing security.

LLM attack differentiation

From an AI view:

- LLM01
- LLM03
- LLM06
- LLM08
- LLM09

OWASP Top 10 for Large Language Model Applications version 1.1

LLM01: Prompt Injection

Manipulating LLMs via crafted inputs can lead to unauthorized access, data breaches, and compromised decision-making.

LLM02: Insecure Output Handling

Neglecting to validate LLM outputs may lead to downstream security exploits, including code execution that compromises systems and exposes data.

LLM03: Training Data Poisoning

Tampered training data can impair LLM models leading to responses that may compromise security, accuracy, or ethical behavior.

LLM04: Model Denial of Service

Overloading LLMs with resource-heavy operations can cause service disruptions and increased costs.

LLM05: Supply Chain Vulnerabilities

Depending upon compromised components, services or datasets undermine system integrity, causing data breaches and system failures.

LLM06: Sensitive Information Disclosure

Failure to protect against disclosure of sensitive information in LLM outputs can result in legal consequences or a loss of competitive advantage.

LLM07: Insecure Plugin Design

LLM plugins processing untrusted inputs and having insufficient access control risk severe exploits like remote code execution.

LLM08: Excessive Agency

Granting LLMs unchecked autonomy to take action can lead to unintended consequences, jeopardizing reliability, privacy, and trust.

LLM09: Overreliance

Failing to critically assess LLM outputs can lead to compromised decision making, security vulnerabilities, and legal liabilities.

LLM10: Model Theft

Unauthorized access to proprietary large language models risks theft, competitive advantage, and dissemination of sensitive information.

OWASP Top 10 for Large Language Model Applications
https://owasp.org/www-project-top-10-for-large-language-model-applications/

ERNW
providing security.

Die Probleme

- Fehlende Trennung von Daten und Anweisungen
- Der Suchraum ist riesig
- Die Modelle sind non-deterministisch
- Die Modelle sind darauf ausgelegt hilfreich zu sein

ERNW
providing security.

Fehlende Trennung von Daten und Anweisungen

- Es gibt keine echte Trennung zwischen…
  - Den Anweisungen an die AI (Instructions)
  - Dem User Prompt
  - Dem Output der AI

  Aus Sicht des LLMs ist alles gemeinsam ein Input aus dem ein Output generiert wird

  -> Das macht es extrem schwer einzugrenzen, was passiert

ERNW
providing security.

Extrem großer Suchraum

Klassischerweise:
Beispiel: Ich suche ein Cross-Site-Scripting in einer Webseite
-> Domänenspezifisch, eingrenzbar, endlich

Trotzdem sehr komplex und bis heute ein Problem in Anwendungen!

ERNW
providing security.

Extrem großer Suchraum

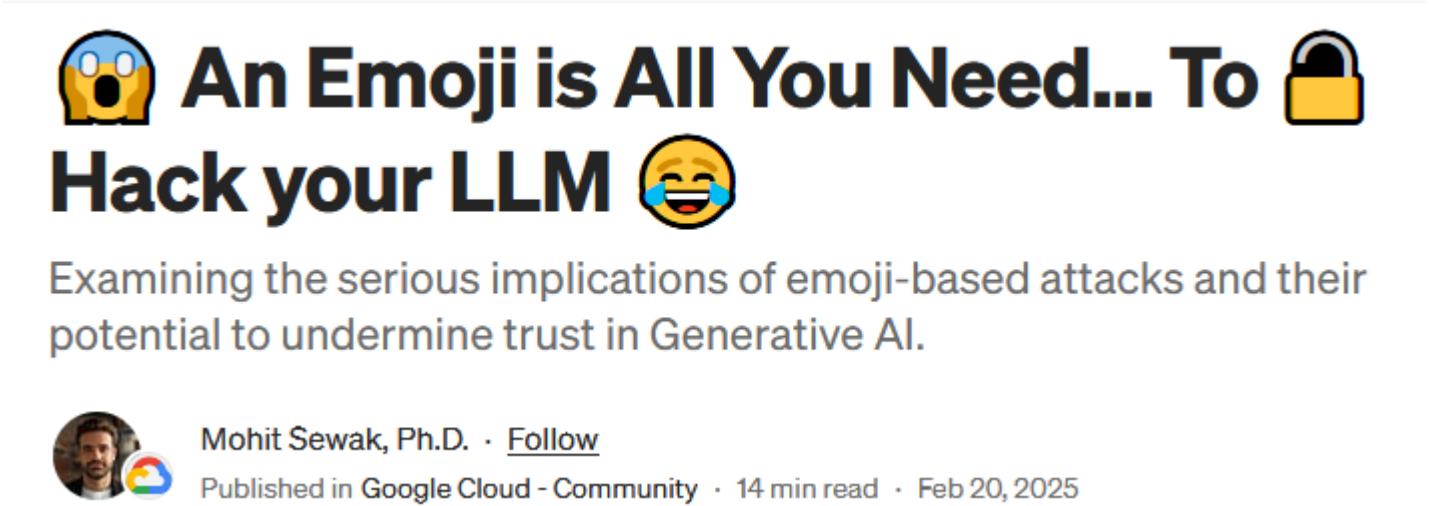
Jetzt:

Der Suchraum für Angriffe ist die…

- Gesamtheit der natürlichen Sprache
- und Emojis
- und merkwürdige Rollenspiele
- Und weitere unbekannte Vektoren

ERNW
providing security.

Extrem großer Suchraum



😱 An Emoji is All You Need... To 🔒 Hack your LLM 😂

Examining the serious implications of emoji-based attacks and their potential to undermine trust in Generative AI.

Mohit Sewak, Ph.D. · Follow

Published in Google Cloud - Community · 14 min read · Feb 20, 2025

Quelle: https://medium.com/google-cloud/emoji-jailbreaks-b3b5b295f38b

ERNW
providing security.

Fehlender Determinismus

Die AI ist darauf trainiert hilfreich, spontan und kreativ zu sein!

-> Gleiche Eingaben führen zu unterschiedlichen Ergebnissen

ERNW
providing security.

mina fahmi @minafahmi_ · 29. Dez. 2022

GPT-3.5 (text-davinci-003) is non-deterministic when temp=0, Top P=1.

Observed within multi-step CoT prompts.

This seems like a regression. Have others observed this?
@OpenAI @goodside

9 6 68 25.191

Boris Power
@BorisMPower

This happens with all the models in our API when there's a tiny difference (<1%) in probability between the two top tokens, due to non determinism.

Once you get one different token then the completions might start to diverge more

Post übersetzen

6:57 nachm. · 29. Dez. 2022 · **37.189** Mal angezeigt

ERNW
providing security.

Die Probleme in Kombination

- Fehlende Trennung von Daten und Anweisungen
- Der Suchraum ist riesig
- Die Modelle sind non-deterministisch

Zusammen ist das eine Katastrophe aus IT-Sicherheitsperspektive!

ERNW
providing security.

Ein Paradigmenwechsel in der IT Security?

- Der traditionelle Umgang mit Schwachstellen hat hier teils keine Bedeutung mehr
  - Angriffe sind nicht reproduzierbar -> Test sind nicht reproduzierbar
  - Maßnahmen schwer implementierbar (AI red teaming / hardening)
  - Metriken lassen sich nur schwer festlegen und nicht wirklich beweise

- Aber… klassische Security Maßnahmen haben noch Gültigkeit!

ERNW
providing security.

Fazit - Was muss ich tun

- Assume Breach
  - Im Moment ist es besser davon auszugehen, dass die AI vom Angreifer kontrolliert wird
- Trust Boundaries festlegen
  - Handlungsspielraum eingrenzen
- Infrastruktur absichern
  - Klassische Security Maßnahmen für alle umliegenden Systeme
  - Mehrere Verteidigungslinien

AI im Unternehmen

ERNW
providing security.

AI im Unternehmen

- Wo setzt ihr Unternehmen AI und was bedeutet das?

How AI is used in business?

From sources across the web

- Improved customer experience
- Personalisation
- Fraud detection
- Accounting
- Demand forecasting
- Marketing
- Cost reduction
- Risk management
- Process automation
- Customer service chatbots
- Increased efficiency
- Better decisions
- AI accounting
- Analytics
- Human Resources
- Sentiment analysis
- Automating manual tasks
- Data analysis and insights
- Predictive analytics
- Customer service operations
- AI for customer service
- Content management
- Marketing and Sales
- Supply chain optimization

Show less

Feedback

ERNW
providing security.

AI im Unternehmen – potenziell überall

- Aber was sind die Dinge, auf die ich wirklich achten muss?

ERNW
providing security.

AI im Unternehmen

Aus unserer Sicht gibt es drei wichtige Themen

- Angriffe gegen die eingesetzte AI (siehe oben)
- Nachvollziehbarkeit (siehe oben, Determinismus)
- Datensicherheit / Datenhoheit / Datenschutz

Datensicherheit

ERNW
providing security.

Is 2025 the Year AI Runs Out of Data?

Brian Bates · Follow

4 min read · Jan 30, 2025

Quelle: https://medium.com/@bates.bp/is-2025-the-year-ai-runs-out-of-data-e5298131566c

ERNW
providing security.

AI companies are desperate for data and they'll go to any length to find it

Enrique Dans · Follow

Published in Enrique Dans · 2 min read · Apr 7, 2024

Quelle: https://medium.com/enrique-dans/ai-companies-are-desperate-for-data-and-theyll-go-to-any-length-to-find-it-e0a7928fcee3

ERNW
providing security.


The New York Times

The Times Sues OpenAI and Microsoft Over A.I. Use of Copyrighted Work

Millions of articles from The New York Times were used to train chatbots that now compete with it, the lawsuit said.

Quelle: https://www.nytimes.com/2023/12/27/business/media/new-york-times-open-ai-microsoft-lawsuit.html

ERNW
providing security.

Datensicherheit – Wir stellen fest

- Den AI-Firmen gehen die Trainingsdaten aus
- Die schon erschöpften Datenquellen sind bereits jetzt fragwürdig

- Sollte ich meine Daten dort hin geben?
- Kann ich nachvollziehen, was damit passiert?
- Kann ich beantragen, dass meine Daten gelöscht werden?

ERNW
providing security.

Datensicherheit – Was kann ich tun?

- Daten anonymisieren
- Metadaten beachten
- Modelle lokal aufsetzen
- Datensparsamkeit!

ERNW
providing security.

AI und Cybercrime

- Wie groß ist das Problem?

ERNW
providing security.

Taken from:

Generative AI Misuse: A Taxonomy of Tactic and Insights from Real-World Data

https://arxiv.org/pdf/2406.13843

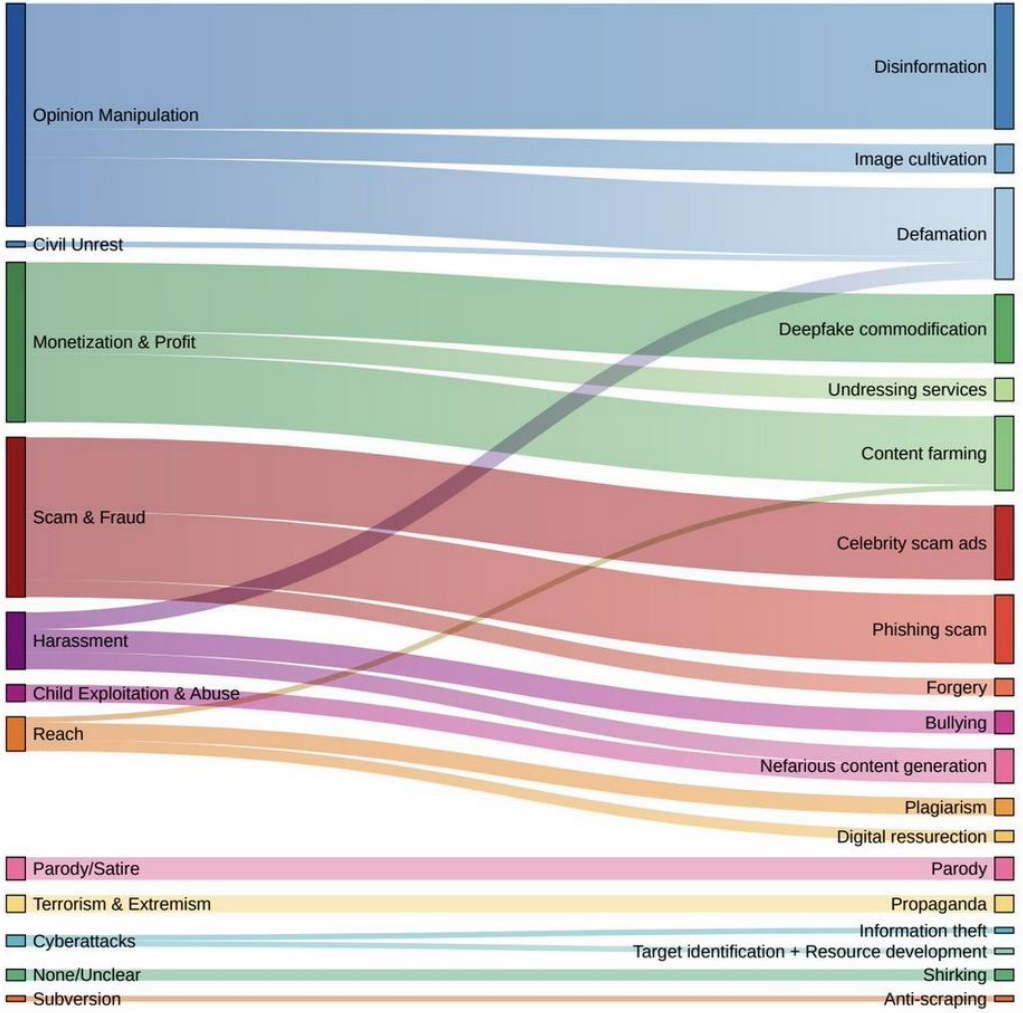


Figure 2 | Top strategies associated with each misuse goal.

ERNW
providing security.

Fazit

- Man muss sich der Schwächen bewusst sein
  - Kompromittierung kann nicht ausgeschlossen werden
  - AI ist nicht verlässlich/deterministisch/nachvollziehbar/testbar
- Man muss die klassischen Maßnahmen ergreifen
  - Datensparsamkeit
  - Need to know Prinzip
  - Minimierung von Privilegien
  - Assume Breach und Multiple Lines of Defense

ERNW
providing security.

You

what is a good way to end a presentation?

Choose an ending that aligns with the tone and content of your presentation while leaving your audience feeling informed, inspired, and motivated to take action.

ChatGPT

Ending a presentation effectively is crucial for leaving a lasting impression on your audience. Here are some tips for a strong conclusion:

1. **Summarize Key Points:** Recap the main points of your presentation to reinforce the key messages you want your audience to remember.
2. **Offer a Call to Action:** Encourage your audience to take specific actions based on the information you've presented. This could be anything from implementing a new strategy to conducting further research.
3. **End with a Powerful Quote:** Close with a relevant and impactful quote that resonates with your presentation's theme or message.
4. **Pose a Thought-Provoking Question:** Leave your audience thinking by asking a question that challenges them to consider new perspectives or ideas related to your topic.
5. **Share a Personal Anecdote:** Conclude with a brief personal story or anecdote that ties back to your presentation's content, adding a human touch and making a deeper connection with your audience.

   Provide Additional Resources: Offer additional reading materials, links, or resources for audience members who want to delve deeper into the topic.
7. **Express Gratitude:** Thank your audience for their time and attention. Expressing gratitude shows appreciation and leaves a positive impression.
8. **End with Impact:** Finish with a memorable statement or image that leaves a lasting impression and reinforces your main message.

Choose an ending that aligns with the tone and content of your presentation while leaving your audience feeling informed, inspired, and motivated to take action.

ERNW
providing security.

Zeit für Fragen?

Danke für die Aufmerksamkeit!

fgrunow@ernw.de
hmohr@ernw.de

@0x79

www.ernw.de

www.insinuator.net

ERNW
providing security.

Make the world a safer place!